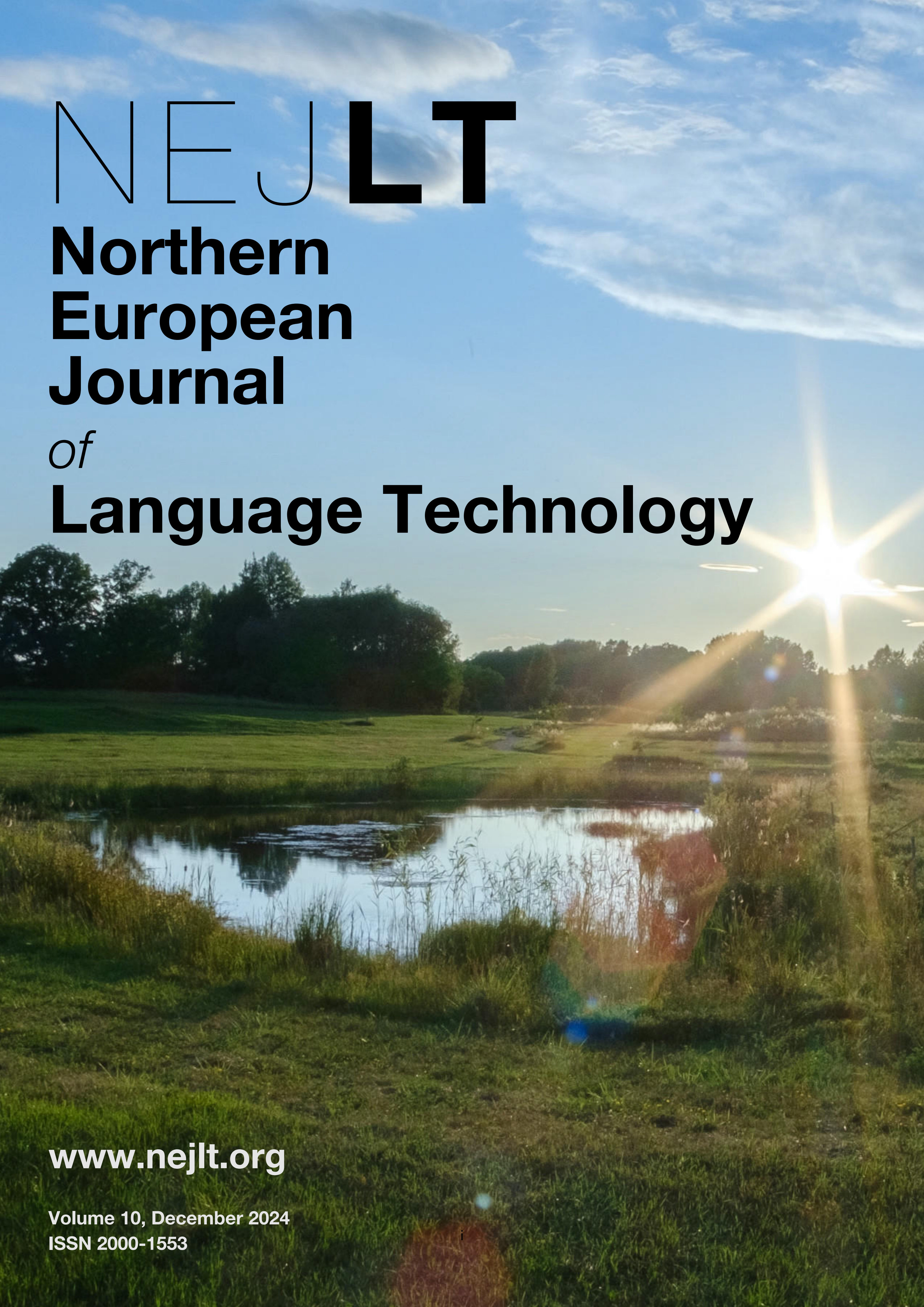


NEJLT

Northern European Journal *of* Language Technology

www.nejlt.org

Volume 10, December 2024
ISSN 2000-1553

Northern European Journal of Language Technology (NEJLT)

Volume 10

December 2024

Editor-in-chief: Marcel Bollmann, editor@nejlt.org

Volume 10 published in 2024 by Linköping University Electronic Press.

LiU Electronic Press
Linköping University
581 83 Linköping
Sweden
ep@ep.liu.se

ISSN 2000-1553

DOI 10.3384/nejlt.2000-1533.10.1

Cover image: © 2023 Marcel Bollmann, https://flic.kr/p/2pS2iVm

The Northern European Journal of Language Technology (NEJLT) is overseen by the Northern European Association for Language Technology (NEALT), a scientific non-profit association.

Editorial Board 2024

Foreword to NEJLT Volume 10

Marcel Bollmann, Linköping University

December 2024

This volume of NEJLT, the Northern European Journal of Language Technology, is the first I have the honour of publishing after taking over the role of editor-in-chief. Leon Derczynski, my predecessor, transformed NEJLT into a modern, global NLP journal with a focus on a fast reviewing turnaround—continuing this ambitious project is both exciting and challenging.

In 2024, NEJLT received twelve submissions; of those that were admitted to the reviewing stage and received a decision, the average time until the first decision was 58 days. Three of those submissions were accepted for publication in this volume; three more submissions that are published here were first submitted in 2023. I am pleased to see a wide geographic distribution of authors on these accepted papers, representing Australia, Austria, Canada, Denmark, Malaysia, Slovakia, Spain, Sweden, the United Arab Emirates, the United Kingdom, the USA, as well as Arabic and African NLP initiatives. This illustrates NEJLT's ambition to be a global journal, and I would like to thank each of these authors for placing their trust in the journal.

As a consequence of the editor-in-chief transition and unclear internal processes, some of the submissions from 2023 unfortunately experienced unacceptably long delays, for which I am sorry. I have done a lot of "invisible" work during this year, clarifying and improving the journal's internal processes, documentation, and tools. One "visible" change, resulting from the adoption of ACL tools and pipelines, is that this issue is the first with more detailed frontmatter and explicit page numbers. I hope that all of this will put NEJLT into a position to continue growing and to operate (more) effectively in the coming years, and also provide faster and more consistent reviewing timelines.

Finally, I would like to thank every editorial board member and every reviewer for contributing their time and efforts towards NEJLT—none of this would be possible without you! I am looking forward to the next year, and hope that it will see many more interesting submissions to the journal.

Table of Contents

Letters

Articles

Efficient Structured Prediction with Transformer Encoders

Ali Basirat, Centre for Language Technology, University of Copenhagen alib@hum.ku.dk

Abstract Finetuning is a useful method for adapting Transformer-based text encoders to new tasks but can be computationally expensive for structured prediction tasks that require tuning at the token level. Furthermore, finetuning is inherently inefficient in updating all base model parameters, which prevents parameter sharing across tasks. To address these issues, we propose a method for efficient task adaptation of frozen Transformer encoders based on the local contribution of their intermediate layers to token representations. Our adapter uses a novel attention mechanism to aggregate intermediate layers and tailor the resulting representations to a target task. Experiments on several structured prediction tasks demonstrate that our method outperforms previous approaches, retaining over 99% of the finetuning performance at a fraction of the training cost. Our proposed method offers an efficient solution for adapting frozen Transformer encoders to new tasks, improving performance and enabling parameter sharing across different tasks.

1 Introduction

The text encoder models evolved from the Transformer architecture (Vaswani et al., 2017) have extensively influenced natural language processing. The standard workflow of these models is based on transfer learning from a model pre-trained on vast amounts of text to a target task. Finetuning is a commonly used technique adjusting the parameters of a pre-trained encoder to a target task using the standard backpropagation algorithm. Despite its simplicity and tremendous success, finetuning can be computationally expensive, particularly for structured prediction tasks in which the parameters are updated at the token level, as opposed to document classification tasks in which the parameters are updated for each document.[1]

In addition, due to in-place parameter updates, finetuning limits the reusability of Transformer encoders, particularly in cloud environments where resources are shared between users. Additionally, finetuning a Transformer encoder for a specific task or a limited number of tasks does not necessarily perform well on other tasks that are not similar to the target task. This is because of catastrophic forgetting, which reduces the generalizability of neural network performance on out-of-domain data (McCloskey and Cohen, 1989). Consequently, a finetuned Transformer encoder becomes a massive computational block specified for a target task, which limits the scalability, modularity, and compositionality of the base encoder (Pfeiffer et al., 2020).

Recent attempts to resolve the shortcomings of finetuning are based on the adapter mechanism (Houlsby et al., 2019) that injects learning blocks into a frozen encoder to facilitate knowledge transfer from pre-training to target tasks (Pfeiffer et al., 2021; Stickland and Murray, 2019; Guo et al., 2021; Hu et al., 2022). While this approach can effectively replicate the finetuning performance (Hu et al., 2022), it still requires careful considerations to balance the encoder's sharability and inference efficiency in a cloud environment, as it tends to sacrifice inference efficiency for sharability, and vice versa, as we will empirically show in this study.

We adopt a different strategy based on the early studies of Peters et al. (2018); Kondratyuk and Straka (2019); Hao et al. (2020) for Transformer encoder adaptation. In contrast to the adapter approach (Houlsby et al., 2019), which adds trainable parameters to the encoder, this method pipes the encoder into an aggregation block that adapts the representations obtained from the encoder's intermediate layers to a target task through linear interpolation. Accordingly, this approach does not necessitate changing the base encoder architecture, making it easier to be shared when compared to the adapter solution.

Although the layer aggregation approach is easy to implement, it requires further consideration when applied to structured prediction. The primary reason for this is that the method assumes that the layers' linear weights are a function of the target task solely and independent of input tokens. This means that a layer weight remains constant for all tokens during inference.

[1] For example, finetuning a BERT model can take 25 days for parsing (Kondratyuk and Straka, 2019) and 2 days for relation extraction (Huguet Cabot and Navigli, 2021).

While this assumption might be acceptable for an encoder that is *finetuned* for *document classification*, as (1) finetuning allows for training the substantial parametric capacity of the original encoder, and (2) classification often relies on a single token vector that represents the entire document, it may not hold true for a frozen model employed for structured prediction. In such cases, it is necessary to model the word dynamics based on the available intermediate representations because the model parameters remain fixed throughout training. Furthermore, Peters et al. (2019) suggest that the linear combination of intermediate representations in a frozen model can match the performance of a finetuned model only if the pre-training and target tasks are sufficiently similar. This implies the need for additional customization of the aggregated representations to account for any disparities between the pre-training and target tasks.

Our paper presents an encoder adaptation model that effectively combines the efficiency of a frozen model with the effectiveness of a finetuned model by addressing the weaknesses of the linear aggregation method. Our approach includes two key mechanisms: an aggregation block and a tailoring block. The aggregation block models the dynamics of words by utilizing the intermediate representations of the encoder and introduces an attention mechanism that trains token representations based on both the target task and the local contribution of intermediate layers to tokens. The tailoring block reduces the impact of dissimilarity between the pretraining and target tasks by refining the aggregated representation, thus allowing for more effective knowledge transfer from pre-training to the target task.

We are not the first to propose the intermediate layer aggregation at the token level. Cao et al. (2022) have also explored a similar approach. They train the token-layer attention weights based on a task-specific query vector used to measure the similarity between intermediate representations locally. Nonetheless, our experiments demonstrate the practical benefits of our approach in downstream tasks in structured prediction and document classification.

We evaluated the effectiveness of our adaptation mechanism on two major classes of Transformer encoders: BERT-based (Devlin et al., 2019) and GPT-based (Radford et al., 2019) models. Our evaluation of the proposed adaptation mechanism builds upon the promising results of ablation studies of the two key blocks. Our experimental results confirm that the adaptation mechanism can perform as well as finetuning while being approximately 13 times more efficient in training time and consuming only 0.3% of the memory required for finetuning, with almost no harm to the inference efficiency. Compared to other approaches, our technique performs significantly better on the majority of structured prediction tasks and remains on par with the best-performing models for document classification.

2 Related Work

The initial investigation of BERT showed that it captures a rich hierarchy of linguistic knowledge in its intermediate layers (Tenney et al., 2019b,a; Jawahar et al., 2019; Lin et al., 2019; de Vries et al., 2020); This leads to the effective use of BERT by linearly aggregating the middle layers based on a target task (e.g. sentiment analysis (Horne et al., 2020; Xiao et al., 2021), morphosyntactic parsing (Kondratyuk and Straka, 2019), gender debiasing and coreference resolution (Abzaliev, 2019), and cross-lingual transfer learning (Chen et al., 2022)). Building on this approach, Cao et al. (2022) expand the task-oriented layer aggregation to encompass both token and task aspects. This extension is achieved through the introduction of an attention fusion model that leverages the local features of a token. We extend Cao et al. (2022)'s method by introducing an attention mechanism that incorporate global views of intermediate representations.

Other techniques that combine intermediate representations include Su and Cheng (2019), who apply the squeeze and excitation technique (Hu et al., 2018), and Yang and Zhao (2019), who use a bidirectional GRU layer to calculate the linear weights of the intermediate representations. From an architectural point of view, our adapter mechanism is based on a dynamic aggregation of the intermediate representations and preserves the parallel encoding functionality of the original encoder. This is in contrast to the linear method of Kondratyuk and Straka (2019), which is based on a static weighting of the intermediate representations, as well as the methods of Yang and Zhao (2019), Horne et al. (2020) and Xiao et al. (2021), which add overhead sequential units to the encoder, hampering parallel computing.

When it comes to improving efficiency, the literature has proposed strategies such as knowledge distillation (Hinton et al., 2015), attention pruning (Michel et al., 2019), model quantization (Zafrir et al., 2019), low-rank adaptation (Hu et al., 2022), shallow finetuning (Ben Zaken et al., 2022) and prompt tuning (Li and Liang, 2021). Our method aligns with the adaptation techniques category, augmenting a frozen model with a few learning blocks to facilitate knowledge transfer to a target task. A related approach by Houlsby et al. (2019) injects adapter layers into a frozen BERT model to enable model sharing in a cloud environment, specifically for efficient sequential multitask learning. Pfeiffer et al. (2021) address the catastrophic forgetting issue and balancing of different tasks in Houlsby et al. (2019)'s approach. Additionally, Stickland and Murray (2019) enhance the efficiency of the adaptation technique by incorporating a low-rank approximation of the model's key operations. In connection to this work, Hu et al. (2022) introduce an efficient method centered around the low-rank factorization of the expected changes in attention matrices in a transformer encoder.

Our approach diverges from that of Houlsby et al. (2019), Pfeiffer et al. (2021), and Hu et al. (2022) in several ways. In terms of model integration, we envelop the adapter around a pre-trained encoder model, unlike the strategies employed by Houlsby et al. (2019), Pfeiffer et al. (2021), and Hu et al. (2022), who embed the adapter blocks within the original encoder architecture. This distinction allows us to exclude the encoder during training, resulting in a substantial enhancement of training efficiency and facilitating model sharing. In addition, the number of trainable parameters in the models of Houlsby et al. (2019) and Hu et al. (2022) scales with the number of intermediate layers of the original encoder, as opposed to our model in which the number of trainable parameters is almost independent of the number of middle layers. Moreover, our encoder adaptation mechanism uses a dedicated tailoring block to explicitly address the differences between the pre-training and target tasks.

3 Encoder Adaptation

An encoder $\mathcal{B}$ consisting of $l-1$ intermediate Transformer layers plus one input embedding layer transforms an input document $s=(t_1,\ldots,t_n)$ into a three-dimensional tensor:

$$\mathcal{B}: V^n \longrightarrow \mathbb{R}^{l\times n\times d}$$

where V is a vocabulary, d is the number of encoder dimensions, and l is the number of intermediate layers. The output tensor $B=\mathcal{B}(s)$ has three views corresponding to its dimensions: A layer view $B_{i,:,:}$ is an $n\times d$ matrix sliced along the layer dimension of B at index i representing a state matrix for the ith layer. A token view $B_{:,j,:}$ is an $l\times d$ matrix sliced along the token dimension of B at index j representing the token at position j. Finally, a feature view $B_{:,:,k}$ is an $l\times n$ matrix sliced along the third dimension of B at index k representing the kth embedding sub-space of encoder feature space.

An adapter function takes an encoding tensor B and merges its layer views (i.e.,) into a matrix:

$$\mathcal{A}: \mathbb{R}^{l\times n\times d} \longrightarrow \mathbb{R}^{n\times d}$$

The resulting matrix includes distilled information of the intermediate representations adapted to a target task. We propose a parametric adaptation function consisting of two blocks: an *aggregation block* that merges the layer views and a *tailoring block* that adjusts the aggregated representations to a target task. The final adapted output is constructed by a residual connection, enabling the model to control the tailoring influence on the aggregation. The architecture is shown in Figure 1.

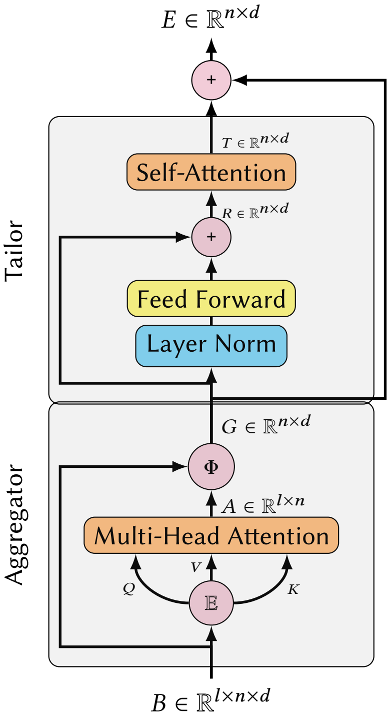

$E \in \mathbb{R}^{n\times d}$
+
$T \in \mathbb{R}^{n\times d}$
Self-Attention
$R \in \mathbb{R}^{n\times d}$
+
Tailor
Feed Forward
Layer Norm
$G \in \mathbb{R}^{n\times d}$
Φ
Aggregator
$A \in \mathbb{R}^{l\times n}$
Multi-Head Attention
V
Q
K
$\mathbb{E}$
$B \in \mathbb{R}^{l\times n\times d}$

Figure 1: The proposed adapter architecture. The input tensor B contains the intermediate representations taken from a Transformer-based encoder for a sequence of tokens, and the output is a token embedding matrix.

3.1 Aggregation

The aggregation block takes the input tensor and merges the layer views. It uses a multi-head attention layer (Vaswani et al., 2017) to calculate the attention weights between pairs of layers and tokens based on the global views of B along with the layer and token dimensions. It then pools the feature vectors that the layers produce for every token based on the attention weights of the token and layers.

Breaking it down step by step, the input tensor B is first passed through the block $\mathbb{E}$ that calculates the attention's query, key, and value matrices. The query matrix Q is a linear projection of the mean matrix $\mathbb{E}_l = \frac{1}{l}\sum B_{i,:,:}$.[2] More formally,

$$Q = \mathbb{E}_l W_Q + b_Q\,,$$

[2] Our preliminary results with uniform and weighted averaging shows that both models perform equally on our development set. Therefore, we select uniform averaging because of its simplicity.

where W_Q is a $d \times k$ trainable matrix, and b_Q is a k-dimensional bias vector with $k \ll d$. Similarly, the key and value matrices are based on a linear projection of $\mathbb{E}_t = \frac{1}{n}\sum B_{:,j,:}$:

$$K = \mathbb{E}_t W_K + b_K \qquad \text{and} \qquad V = K\,,$$

where the $d \times k$ matrix W_K and the k-dimensional bias vector b_K ($k \ll d$) are learnable parameters.[3] We refer to the parameter k as the *attention dimensionality* of the adapter.

Next, the Multi-Head Attention layer constructs an $l \times n$ matrix A whose columns define weight distributions over the layer views for each token. Intuitively, the attention value $A_{i,j}$ indicates the importance of layer view $B_{i,:,:}$ in the representation of the token t_j. Finally, the aggregated representation for a token at position j is calculated in the computational block Φ based on the weighted sum of the corresponding intermediate representations:

$$G_{j,:} = \sum_{i=1}^{l} A_{i,j} B_{i,j,:}$$

This sum pooling is equivalent to $\text{diag}(A^T B)$ where the diag operator is applied on the first and second dimensions of the product $A^T B$.

3.2 Tailoring

The tailor block further specifies the aggregated token vectors to the target task. Inspired by the Transformer architecture (Vaswani et al., 2017), the tailor block adopts residual learning (He et al., 2016) with a residual mapping consisting of a layer normalization (Ba et al., 2016) followed by a position-wise feed-forward network:

$$R = G + \text{ReLU}(\text{Dropout}(L(\text{Norm}(G))))\,,$$

where L is a linear layer that adjusts the aggregated token vectors to the target task. Following Xiong et al. (2020), we use pre-layer normalization in which the normalization layer is placed before the feed-forward network.[4]

Finally, the self-attention layer compensates for the lack of trainable parameters to model task-specific dependencies between tokens. We calculate a self-attention matrix based on a linear transformation of the input vectors:

$$V = \text{Dropout}(R) W_T + b_T \qquad Q = K = V$$

where W_T is a $d \times k$ trainable matrix, and b_T is a k-dimensional bias vector. We then utilize a MultiheadAttention layer (Vaswani et al., 2017) to construct an $n \times n$ attention matrix M based on the query (Q), key (K), and value (V) matrices. The reason for performing the linear transformation on R is to reduce the dimensionality of the token vectors, which determine the size of the learning parameters in the MultiheadAttention block. To calculate the tailored matrix T of size $n \times d$, we multiply the attention matrix by the input matrix R:

$$T = M \times R$$

The rows of T are the sum of the token representations in R weighted by their attention scores.

4 Experiments

We study the adapter performance on downstream tasks and investigate the contribution of the aggregation and tailoring blocks to the performance gain. The experiments are based on the cased versions of BERT-Large (Devlin et al., 2019) and ROBERTA-Large (Liu et al., 2020) models as representatives of encoder-only models, and different variants of the GPT2, including GPT2-Small, Medium, and Large, as representatives of decoder-only models. All models are provided by Huggingface.

The test benchmark includes the two major types of classification tasks in NLP, including structured prediction and document classification. The evaluation benchmark for structured prediction includes tasks defined on the following datasets:

- CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003): part-of-speech tagging (POS), chunking (CHK) and named-entity recognition (NER)
- Universal Dependencies (Nivre et al., 2017) (English EWT, v2.3): part-of-speech tagging (UPOS and XPOS), dependency label prediction (DEPREL), and dependency parsing (LAS)
- English Penn Treebank (Marcus et al., 1993) converted to Stanford Dependencies: dependency parsing (LAS)
- WEBNLG (Gardent et al., 2017): named-entity recognition (NER) and relation extraction (RE)

We use the standard data splits for training, validation, and testing the models. The reason for incorporating multiple datasets is to demonstrate the method's robustness not only across tasks but also in varied data and domains. The document classification benchmark is based on the GLUE tasks (Wang et al., 2018) including grammaticality (CoLA), sentiment textual similarity

[3] We set V equal to K based on our preliminary experiments showing no significant difference in the results with and without a dedicated transformation of the values.

[4] This is in contrast to post-layer normalization, which is used in the original Transformer architecture (Vaswani et al., 2017) and the BERT implementation of Devlin et al. (2019).

(STS-B), semantic equivalence (MRPC), textual entailment recognition (RTE), and sentiment analysis (SST-2). For these tasks, we train and validate models on their training data, and keep the benchmark's validation data for final model testing.

We set the attention dimension k to 16 and the number of attention heads to 2 in both attention layers in the aggregator and adapter blocks. This decision is based on our preliminary results on the CoNLL development data. For each task, we train five classifiers with different random seeds and report the average results on the test sets. As our optimizer, we use Adam with a 1cycle learning rate scheduler with cosine annealing. Our implementations are based on PyTorch (Paszke et al., 2019) and the experiments are carried out on an NVIDIA A100 40GB Tensor Core GPU. An implementation of the encoder model is available here `https://github.com/abasirat/llm-adapter`.

5 Performance

This section summarizes the results collected from the structured prediction and document classification tasks. We compare our adapter performance, named Adapted, with other techniques, namely:

- Frozen: returns the output of the last layer of a frozen encoder (i.e., $\mathcal{A}(B) = B_{l,:,:}$).
- Linear: returns a scaled linear aggregation of intermediate representations of a frozen model, i.e., $\mathcal{A}(B) = c\sum_{i=1}^{l} w_i B_{i,:,:}$ where $w_0, \ldots, w_l = \text{Softmax}(\alpha_0, \ldots, \alpha_l)$ (Kondratyuk and Straka, 2019). We train the parameters α_i and c together with other trainable parameters of each task's classifier.
- Fusion: returns a linear aggregation of the intermediate representations of a frozen model for each token and task, i.e., $\mathcal{A}(B) = \sum_{i=1}^{l}(A \odot B)_{i,:,:}$ where A is an $l \times n$ attention matrix and $A \odot B$ is the Hadamard product between A and every feature view of B. The attention matrix A is based on a d-dimensional task-specific attention vector Q learned during training, i.e., $A_{i,j} = \frac{\exp(Q \times B_{i,j,:})}{\sum_{k=1}^{l} \exp(Q \times B_{k,j,:})}$ (Cao et al., 2022).
- Lora: returns the output of the last layer of an encoder adapted to a target task based on the Lora factorization technique of Hu et al. (2022). Lora is an efficient technique that improves state-of-the-art on most downstream tasks. It extends the adaptation technique of Houlsby et al. (2019) using a simplified version of the method of Aghajanyan et al. (2020). We train Lora with its recommended setting.
- Finetune: returns the output of the last layer of a fine-tuned encoder model (Devlin et al., 2019).

We freeze the encoder parameters in all of the above techniques except for the Finetune, in which all encoder parameters are updated during training. The outputs taken from these encoders are subsequently employed for the target tasks outlined in Section 4. The following sections provide detailed insights into the modules employed for each task.

5.1 Structured Prediction

In this section, we study the adapter performance on BERT and GPT encoder families. The token classifiers for POS tagging, chunking (CHK), and DEPREL prediction consist of an encoder (e.g., BERT) with a light head block mapping a token vector to a probability distribution over target tags. The head block consists of a dropout layer followed by a dense layer of size $d \times d$ (d is the encoder's dimensionality) with $tanh$ activation connected to another dropout and a dense layer of size $d \times |\text{tag set}|$ with a softmax activation. Following Ács et al. (2021), we classify tokens based on their first subword for semantic tasks such as NER in CoNLL2003 and their last subword for syntactic tasks such as POS and CHK in CoNLL2003 and UPOS, XPOS, and DEPREL in UD. We integrate the encoder adapters into the parser of Dozat and Manning (2017) for the parsing experiments and the joint named entity and relation extraction system of Yan et al. (2021) for NER and RE in WEBNLG.[5]

We set the batch size to 32 sentences for the tasks in CoNLL, UD (excluding dependency parsing), and WEBNLG and 5000 samples for dependency parsing on UD and WSJ. We train the CoNLL and UD taggers for 10 epochs, and the parsers and relation extractors for 100 epochs. We schedule the learning rate using the 1cycle policy (Smith and Topin, 2017) with a cosine annealing strategy. The maximum learning rate for parsing is set to $2e-3$, and for other tasks is selected among $\{1e-5, 1e-4, 1e-3\}$ based on the models' performance on the development sets. We disable the parser's character embedding module to better study the adaptation effect. The other parameters in the parser and relation extractor are set to their default values.

5.1.1 BERT Family

Table 1 summarizes the results obtained from each adaptation technique on the structured prediction tasks. The Adapted models based on our proposed approach perform better than Frozen, Linear, and Fusion models in all tasks,[6] and on par with Lora and Finetune

[5] Our parsing experiments are based on the parser implementation available at `https://github.com/Unipisa/biaffine-parser`

[6] One exception is dependency parsing, in which the linear models perform slightly better than the Adapted models.

models. On average, the Adapted models result in extensive improvements over the baseline frozen models by 3.7%, which is significantly larger than the improvement made by the Linear (2.6%) and Fusion (1.4%) models , and is on a par with Lora (3.8%) and Finetune (3.8%) models (see Figure 2 for detailed information).

A deeper look into the results shows that the absolute improvement made by the Adapted models over the baseline Frozen models, as shown in Figure 2, is higher than that of the other techniques on the syntactic tasks, such as POS tagging and chunking, but slightly lower than Lora on the more semantic tasks NER and RE. This suggests that the Adapted models are more effective than other techniques in encoding the syntactic information. A further detailed study on the contribution of the aggregation and tailoring blocks to syntactic and semantic encoding is presented in Section 6.

Also, we see that Adapted models outperform other models in retaining 99.5% of the BERT's finetuning performance (excluding dependency parsing in which the finetuned models perform significantly lower than the frozen baseline) and level with Lora on the ROBERTA's finetuning performance (i.e., the Adapted and Lora models retain 99.2% and 99.4% of finetuned ROBERTA models, respectively.[7]

5.1.2 GPT Family

GPT models are generative language models known for their strong performance in text generation tasks. However, they are less commonly used for discriminative tasks, included in our test benchmarks. The purpose of our experiments with GPT models is to present empirical evidence on the utility of different adapting techniques for the GPT family. We leverage a GPT model as a feature extractor, bypassing its final generator layer. After tokenizing an input string, we perform a forward pass through the GPT model with the tokenized input and extract the hidden states from all layers. These hidden states serve as the encoder output, denoted as B in our adapter formulation in Section 3, which is subsequently adapted for downstream tasks. To manage computational costs, the experiments focus on a smaller number of tasks and exclude Finetune experiments.

Table 2 summarizes the results collected from GPT models. First, compared to the BERT-based models, the results show that the adapted GPT models perform weaker. This performance drop is expected due to the discriminative nature of the tasks that are not generally considered for GPT models. However, within the adapted GPT models, our adapter performs better than other techniques in all tasks.[8] This contrast with the relative model performance on the BERT family, where the Adapted model performs on par with Lora on most tasks. We are uncertain about the cause of this discrepancy in the better relative performance of the Adapted model on the GPT family compared to the BERT family, and we leave it for further investigation in future research.

5.2 Document Classification

While our adaptation technique primarily emphasizes individual token representations rather than document representation, we still find it important to examine the adapter performance on standard document classification tasks, even though it may not be the optimal use case for our technique. We apply the adaptation technique on a subset of GLUE tasks (Wang et al., 2018) and compare its performance with other methods. The experiments are based on the CoLA, STS-B, RTE, SST-2, MRPC, and RTE tasks. We train five models with different random seeds and report the average classification score on the development sets. We select the learning rate from $\{1e-4, 3e-4, 7e-4\}$ for Frozen, Linear, Fusion, Adapted, and Lora techniques and from $\{1e-5, 2e-5, 3e-5\}$ for Finetune models.

We follow the standard approach for document classification that encodes an input document (e.g., a sentence or a pair of sentences) into a vector representation and then passes it into a header block that maps the vector to a class distribution. The document vector is often a dedicated vector (e.g., [CLS] in BERT) or the mean of the token vectors comprising the input document. In this study, we use the mean vector, which relies on the token representations, to represent a document. The classification head is similar to the header block we use for structured prediction networks. Except for SST-2 classification models, we train other models up to 100 epochs with a learning patience of 20. The SST-2 models are trained for 20 iterations and the learning patience of 10. The batch size in all training setups is 32.

Figure 3 summarizes the results of document classification. On average, the Adapted models (79.6) outperform the Frozen (69.7), Linear (71.3), and Fusion (74.3) models but perform slightly weaker than Lora (81.4) and Finetune (81.7) models on the document classification. The higher performance of the Adapted to the Fusion models indicates that our token-layer attention mechanism based on the global contextual information is more meaningful to the document classification tasks

[7] We exclude dependency parsing from the finetuning analysis because it does not perform as expected. We considered different finetuning strategies (finetuning top $n = 1, \ldots, 4$ layers, finetuning middle layers $(9-17)$, followed by linear aggregation, and finetuning only during the first and second epochs. In all experiments, we could get maximum LAS of 94.4 for BERT and 93.8 for ROBERTA on WSJ, which is still significantly below the frozen baseline.

[8] One exception is Lora used for DEPREL prediction.

BERT-Large (345M)

	CoNLL-2003 (English)			UD (English-EWT)				WSJ	WEBNLG	
	POS	CHK	NER	UPOS	XPOS	DEPREL	LAS	LAS	NER	RE
Frozen	89.4 ±0.06	83.9 ±0.05	79.9 ±0.23	92.7 ±0.06	92.8 ±0.06	80.4 ±0.04	95.04 ±0.53	94.83 ±0.05	96.2 ±0.19	89.8 ±0.25
Linear	93.2 ±0.07	89.6 ±0.04	85.6 ±0.09	96.2 ±0.02	96.1 ±0.04	87.8 ±0.02	96.37 ±0.43	**95.24** ±0.09	96.6 ±0.21	91.4 ±0.47
Fusion	92.5 ±0.30	87.9 ±1.32	87.2 ±0.15	95.9 ±0.07	95.8 ±0.14	87.9 ±0.04	95.76 ±0.51	95.07 ±0.07	96.4 ±0.43	90.2 ±1.02
Adapted	93.6 ±0.04	90.4 ±0.17	89.5 ±0.27	96.4 ±0.07	96.3 ±0.06	91.7 ±0.05	**96.63** ±0.36	95.17 ±0.05	97.3 ±0.08	**92.2** ±0.35
Lora	92.8 ±0.17	89.4 ±0.22	89.8 ±0.22	96.0 ±0.07	96.0 ±0.07	91.6 ±0.42	95.09 ±0.50	94.90 ±0.07	97.4 ±0.07	92.0 ±0.21
Finetune	**93.9** ±0.07	**91.1** ±0.07	**91.3** ±0.13	**96.8** ±0.04	**96.8** ±0.04	**93.6** ±0.07	–	94.30 ±0.21	**97.8** ±0.07	91.4 ±0.10

ROBERTA-Large (345M)

	CoNLL-2003 (English)			UD (English-EWT)				WSJ	WEBNLG	
	POS	CHK	NER	UPOS	XPOS	DEPREL	LAS	LAS	NER	RE
Frozen	91.3 ±0.05	87.8 ±0.05	82.8 ±0.12	94.0 ±0.04	93.8 ±0.03	83.8 ±0.08	96.29 ±0.46	95.40 ±0.13	95.6 ±0.11	89.8 ±0.21
Linear	93.2 ±0.03	89.6 ±0.04	86.1 ±0.10	96.3 ±0.02	96.3 ±0.05	87.5 ±0.06	**96.68** ±0.41	**95.53** ±0.04	96.5 ±0.11	91.4 ±0.34
Fusion	91.2 ±0.93	86.2 ±3.50	85.1 ±2.58	95.2 ±0.31	94.8 ±0.43	85.6 ±1.15	96.14 ±0.43	95.31 ±0.15	94.8 ±1.52	88.1 ±2.02
Adapted	93.6 ±0.03	90.5 ±0.12	89.7 ±0.23	96.6 ±0.09	96.4 ±0.04	91.8 ±0.07	96.25 ±0.32	95.45 ±0.12	96.8 ±0.17	91.9 ±0.48
Lora	93.0 ±0.04	89.7 ±0.10	91.6 ±0.13	97.0 ±0.10	97.0 ±0.07	93.1 ±0.07	96.29 ±0.43	95.39 ±0.09	97.4 ±0.16	**92.1** ±0.26
Finetune	**93.7** ±0.10	**91.1** ±0.17	**92.6** ±0.11	**97.6** ±0.04	**97.5** ±0.04	**94.5** ±0.10	–	93.66 ±0.42	**97.8** ±0.06	91.4 ±0.34

Table 1: Encoder adaptation performance on downstream tasks averaged over five trials with different random seeds. All results are based on the F_1 score on the test sets. Bold: the results of best-performing models. Gray highlights: the best-performing model among the non-finetuned models, i.e., the models that preserve the base encoder frozen during training. The comparisons are based on a two-tailed t-test with p-value<0.05

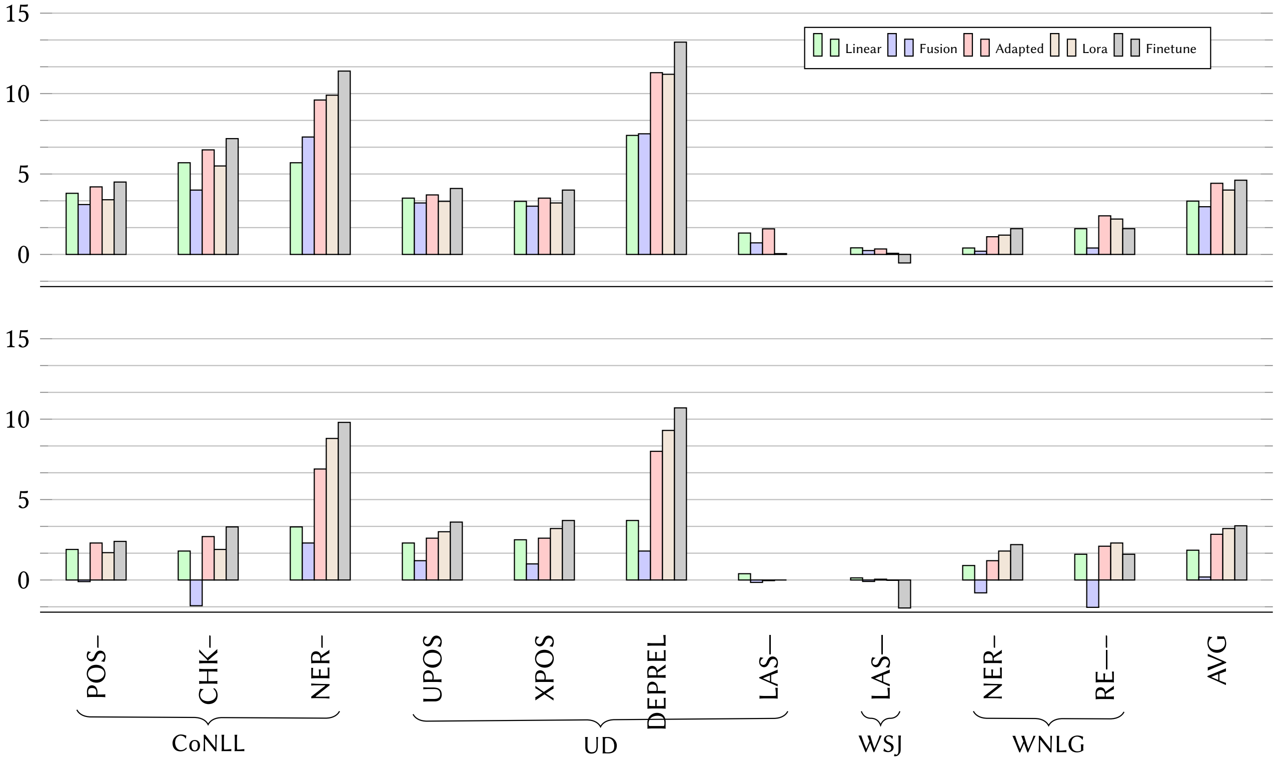


Figure 2: Absolute performance improvement (or degradation) over the frozen baseline. Top: BERT, bottom: ROBERTA.

	CoNLL-2003 (English)			UD (English-EWT)		
	POS	CHK	NER	UPOS	XPOS	DEPREL
			GPT2-small (124M)			
Frozen	90.7 ±0.06	86.5 ±0.07	68.8 ±0.27	92.6 ±0.05	92.8 ±0.04	78.6 ±0.06
Linear	91.4 ±0.29	88.7 ±0.60	63.1 ±0.09	92.9 ±0.22	93.6 ±0.44	80.4 ±0.19
Fusion	89.1 ±0.11	86.1 ±0.30	62.1 ±0.55	89.6 ±1.10	89.4 ±0.84	76.4 ±0.87
Adapted	**93.4** ±0.07	**89.7** ±0.18	**81.0** ±1.50	**94.3** ±0.06	**95.0** ±0.09	**82.8** ±0.11
Lora	92.0 ±0.06	87.6 ±0.07	70.1 ±0.19	92.8 ±0.08	93.4 ±0.05	80.7 ±0.07
			GPT2-medium (355M)			
Frozen	90.3 ±0.07	86.6 ±0.09	71.8 ±0.22	92.7 ±0.05	92.8 ±0.15	78.8 ±0.13
Linear	91.8 ±0.26	88.9 ±0.21	64.2 ±0.13	94.2 ±0.04	94.5 ±0.29	81.4 ±0.56
Fusion	90.0 ±0.32	85.0 ±2.64	70.3 ±4.60	92.7 ±0.20	92.8 ±0.22	78.4 ±0.37
Adapted	**93.5** ±0.07	**89.3** ±0.66	**80.0** ±0.74	**94.3** ±0.22	**95.2** ±0.18	**82.8** ±0.72
Lora	92.7 ±0.06	88.6 ±0.06	73.2 ±0.10	93.5 ±0.03	94.4 ±0.04	**83.3** ±0.04
			GPT2-large (774M)			
Frozen	90.8 ±0.05	86.7 ±0.08	72.5 ±0.16	93.2 ±0.06	93.2 ±0.05	79.5 ±0.10
Linear	93.2 ±0.04	89.4 ±0.15	68.5 ±0.47	94.5 ±0.07	95.2 ±0.02	83.0 ±0.06
Fusion	92.1 ±0.15	87.5 ±0.32	72.8 ±0.48	93.7 ±0.15	94.1 ±0.09	80.8 ±0.10
Adapted	**93.4** ±0.09	**89.6** ±0.10	**83.3** ±0.37	**94.7** ±0.08	**95.4** ±0.12	**83.9** ±0.12
Lora	93.0 ±0.03	89.1 ±0.09	75.7 ±0.40	93.7 ±0.04	94.7 ±0.03	**84.0** ±0.07

Table 2: The performance of the adapter techniques on GPT models. Bold numbers are significantly higher than other results in a column (two-tailed t-test with p-value<0.05).

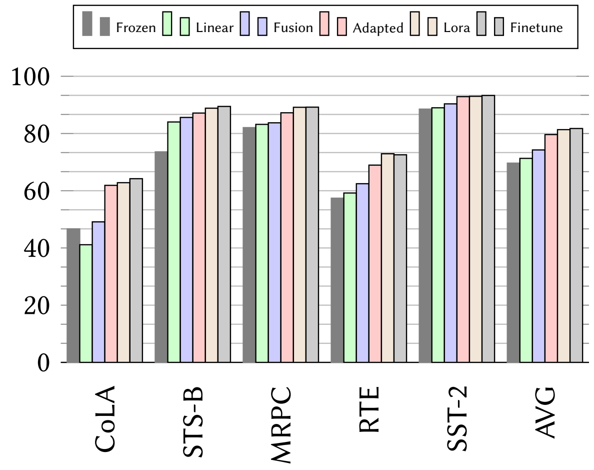


Figure 3: Document classification results based on the BERT-Large model.

than the locally trained attention weights trained of the Fusion models. The Adapted models perform comparably to Lora on the sentiment tasks (STS-B and SST-2) and the grammaticality task CoLA but weaker on other tasks. This observation indicates that tuning the internal attention weights of the encoder, as done by Lora, is more beneficial to the document classification tasks than adapting the intermediate representations, as our Adapted model considers.

5.3 Model Efficiency

The adapting techniques we consider in this study can be classified into two categories regarding their integration into the base model. The first are those techniques that chain a learning block on top of the encoder, and the second are those that are infused into the encoder. The piped models treat the base encoder as a black box and rely only on the intermediate representations produced by the encoder. However, the infused techniques must carefully modify the base architecture and insert their parameters in the encoder. In our test benchmark, the learning blocks of the Linear, Fusion, and Adapted techniques are chained onto the base encoder while the parameters of Lora are infused.

When it comes to the standard model training with backpropagation, the piped techniques are more advantageous since they allow us to perform the forward pass only once, store the intermediate representation (i.e., the layer activations), and reuse them during training. This trick significantly improves training time at the cost of storage, which is much cheaper than the GPU cost required. However, both piped and infused techniques perform the backward pass similarly on all trainable parameters. Therefore, in order to improve training efficiency, we adopt the activation storing trick for Frozen, Linear, Fusion, and Adapted models.

Table 3 summarized the training and inference efficiency of the adapter techniques. The results are averaged over the efficiency statistics of CoNLL tagging models. Training the piped model is significantly more efficient than training the infused models. The Frozen, Linear, and Fusion models are the most efficient techniques in both computational time and trainable parameters. The Adapted models are more than 10× faster than Finetune and Lora models but 2× slower than Frozen, Linear, and Fusion models, which is due to the large number of trainable parameters it uses. The training efficiency gain of the piped models comes from the activation storing trick.

The activation storing trick is also responsible for the shorter training time in piped models compared to their inference time. As mentioned earlier, we forward every training example only once through the base encoder of these models during their first training epoch. Later, we only pass it through the adapter parameters, which are relatively smaller than the encoder. Assuming t_e as the time required for a forward pass through the encoder, t_f, and t_b as the time required for a forward and backward pass through the adapter parameters, then the average training time for a batch of sentences is $t_{\text{tr}} = \frac{t_e + e(t_f + t_b)}{e}$ for e training epochs. This fraction becomes smaller as e increases. However, the inference time for a test example is $t_{\text{inf}} = t_e + t_f$, which is larger than the average training time if the number of training epochs (e) is larger than 1. This is because $t_e + et_b < et_e$ for $e > 1$ given that $t_b \ll t_e$, which implies $t_{\text{tr}} < t_{\text{inf}}$. It is important to note that the training times reported in Table 3 are averaged over the training epochs, and the total training time is longer than the inference time.

The Adapted models perform as efficiently as the other piped models at the inference time while still being more efficient than Lora. Lora's lower inference efficiency is because of the infused attention weights distributed across all encoder layers. Although the added parameters are relatively small, they still cause a computational delay because they are coupled to the encoder's attention weights in each layer and must be processed sequentially in the same order as the encoder's layer. This latency can be improved by merging the infused weights into the encoder's attention parameters. In this case, Lora will be as efficient as the other models but at the cost of losing its reusability because it will become a large model specified for a target task. In contrast, our adaptation mechanism includes a small number of relatively large computational blocks to the encoder, enabling more efficient use of the GPU parallel processing capability.

	Train	Inference	#Params
Frozen	0.004	0.024	0.0
Linear	0.004	0.024	0.0
Fusion	0.004	0.024	0.0
Adapted	0.008	0.025	1.1
Lora	0.095	0.062	0.8
Finetune	0.106	0.024	333.6

Table 3: The model efficiency. Train/Inference: average training/inference time (seconds) for a batch of 32 sentences. The time does not include the tokenization and data loading time, which is independent of the actual training. The training time is averaged over training epochs. #Params: number of trainable parameters in each adaptation mechanism ($\times 10^6$).

6 Ablation Study

This section studies the importance of the aggregation and tailoring blocks to our adapter architecture. Figure 4 represents the improvements made by each block over the baseline Frozen models. A significant part of the improvement in the structured prediction tasks comes from the aggregation that accounts for 85% of the average improvements over frozen BERT models. However, both blocks contribute almost equally to the average improvement in document classification. This indicates the importance of the token-wised layer aggregation for structured prediction tasks that search for the interconnections between tokens. On the other hand, the tailoring block contributes significantly to the document classification tasks whose objectives differ from the encoder's pretraining objective (i.e., masked token prediction).

We also see that the necessity for tailoring in the structured prediction becomes more evident as the task complexity increases. The results suggest that the required information for syntactic tasks such as POS tagging, chunking, and parsing is already available in the intermediate representations and we only need to aggregate the information properly. However, more complex tasks that rely on both syntactic and semantic inference (e.g., DEPREL, NER, and RE) can benefit from both aggregation and tailoring blocks.

Next, we study the importance of the residual connection between the aggregation and tailoring blocks.

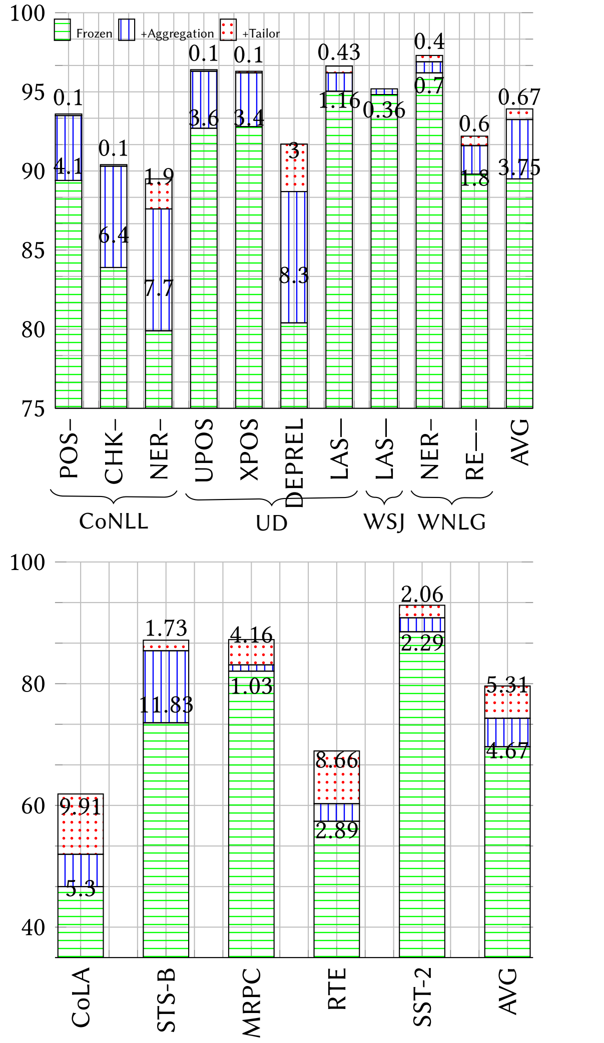


Figure 4: The accumulative contribution of the aggregation and tailoring blocks to the frozen BERT model. The adapted output is taken from the residual connector, adding the aggregator and tailor outputs.

Table 4 summarizes the base results on a subset of the structured prediction tasks. The base results in the Agg column are from the aggregator block. The other columns summarize the improvement or degradation caused by the tailoring block and the residual connection between the two blocks. First, compared with the results reported in Table 1, the aggregator block (see Column Agg) surpasses the Linear and Fusion aggregation by an average score of 0.6 and 1.9, respectively. This empirically supports our assumption about the local contribution of the intermediate representations to tokens within a task, as opposed to the token-independent aggregations of the Linear model. It also shows that our global modeling of the token-layer attention mechanism performs better than the local modeling of the Attention Fusion mechanism. Second, the tailoring block hurts the model performance when piped to the aggregation block without the residual link (see Column +Tailor). The residual information cancels out the tailor negative effect and significantly improves performance. We speculate that this is due to the controlling effect of the residual connection that lets the tailoring block affect the aggregated information only in specific contexts if needed.

	Agg	+Tailor	+Residual	Adapted
POS	93.6	0.0	0.1	0.1
CHK	90.3	−1.0	1.1	0.1
NER	87.4	0.0	1.6	1.6
DEP	95.2	−0.2	0.2	0.0
RE	91.6	−0.3	0.9	0.6
AVG	91.6	−0.3	0.8	0.5

Table 4: The performance improvement (or degradation) after adding the tailoring block and the residual link to the aggregation block. The results are based on the BERT model.

7 Conclusion

We have introduced a task adaptation mechanism for Transformer encoders to address the reusability and efficiency issues of finetuning in structured prediction. The proposed model aggregates the intermediate representations of a frozen encoder based on the input tokens and tailors them to a target task. Empirical results confirmed that the adaptation mechanism improves the training efficiency significantly while being on par with the finetuning performance. Further ablation studies confirmed the importance of both the aggregation and tailoring blocks. In future work, we want to study the adapter performance within different languages and analyze attention weights in more detail across different tasks in multilingual benchmarks.

Acknowledgements

We are grateful to the unknown reviewers who provided invaluable feedback on this work and to Marco Kuhlmann for his interest, contributions, and support for this research. We would also like to express our thanks to the Danish National Life Science Supercomputing Center for granting access to Computerome 2.0 through Project ku-00223 and to the National Supercomputer Centre in Sweden (NSC) for allowing access to Berzelius through Project Berzelius-2021-51. This project was supported by the Excellence Center at

Linköping-Lund in Information Technology (ELLIIT), Project A15.

References

Abzaliev, Artem. 2019. On GAP coreference resolution shared task: Insights from the 3rd place solution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 107–112, Florence, Italy. Association for Computational Linguistics.

Ács, Judit, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Aghajanyan, Armen, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning.

Ba, Lei Jimmy, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Ben Zaken, Elad, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Cao, Jin, Chandana Satya Prakash, and Wael Hamza. 2022. Attention fusion: a light yet efficient late fusion mechanism for task adaptation in NLU. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 857–866, Seattle, United States. Association for Computational Linguistics.

Chen, Beiduo, Wu Guo, Quan Liu, and Kun Tao. 2022. Feature aggregation in zero-shot cross-lingual transfer using multilingual bert.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dozat, Timothy and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Gardent, Claire, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Guo, Demi, Alexander Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4884–4896, Online. Association for Computational Linguistics.

Hao, Yaru, Li Dong, Furu Wei, and Ke Xu. 2020. Investigating learning dynamics of BERT fine-tuning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 87–92, Suzhou, China. Association for Computational Linguistics.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE Computer Society.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Horne, Leo, Matthias Matti, Pouya Pourjafar, and Zuowen Wang. 2020. GRUBERT: A GRU-based method to fuse BERT hidden layers for Twitter sentiment analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 130–138, Suzhou, China. Association for Computational Linguistics.

Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Hu, Edward J, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu

Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hu, Jie, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

Huguet Cabot, Pere-Lluís and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Kondratyuk, Dan and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Li, Xiang Lisa and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Lin, Yongjie, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: a robustly optimized BERT pretraining approach.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

McCloskey, Michael and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Michel, Paul, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nivre, Joakim, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peters, Matthew E., Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 7–14, Florence, Italy. Association for Computational Linguistics.

Pfeiffer, Jonas, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Pfeiffer, Jonas, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1.

Smith, Leslie N. and Nicholay Topin. 2017. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*, abs/1708.07120.

Stickland, Asa Cooper and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

Su, Ta-Chun and Hsiang-Chih Cheng. 2019. Sesame-bert: Attention for anywhere.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR 2019)*.

Tjong Kim Sang, Erik F. and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

de Vries, Wietse, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xiao, Zeguan, Jiarun Wu, Qingliang Chen, and Congjian Deng. 2021. BERT4GCN: Using BERT intermediate layers to augment GCN for aspect-based sentiment classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9193–9200, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiong, Ruibin, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.

Yan, Zhiheng, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 185–197, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang, Junjie and Hai Zhao. 2019. Deepening hidden representations from pre-trained language models.

Zafrir, Ofir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. Q8bert: Quantized 8bit bert. In *Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*, volume 5, pages 36–39.

DANSK: Domain Generalization of Danish Named Entity Recognition

Kenneth Enevoldsen, Aarhus University kenneth.enevoldsen@cas.au.dk

Emil Trenckner Jessen, Aarhus University

Rebekah Baglini, Aarhus University

Abstract Named entity recognition is an important application within Danish NLP, essential within both industry and research. However, Danish NER is inhibited by a lack coverage across domains and entity types. As a consequence, no current models are capable of fine-grained named entity recognition, nor have they been evaluated for potential generalizability issues across datasets and domains. To alleviate these limitations, this paper introduces: 1) DANSK: a named entity dataset providing for high-granularity tagging as well as within-domain evaluation of models across a diverse set of domains; 2) and three generalizable models with fine-grained annotation available in DaCy 2.6.0; and 3) an evaluation of current state-of-the-art models' ability to generalize across domains. The evaluation of existing and new models revealed notable performance discrepancies across domains, which should be addressed within the field. Shortcomings of the annotation quality of the dataset and its impact on model training and evaluation are also discussed. Despite these limitations, we advocate for the use of the new dataset DANSK alongside further work on generalizability within Danish NER.

1 Introduction

Danish **A**nnotations for **N**LP **S**pecific Tas**K**s (**DANSK**) is a new gold-standard dataset for Danish with named entity annotations for 18 distinct classes. The annotated texts are from 25 text sources that span 7 different domains and have been derived from the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021). The dataset is publicly accessible[1] and pre-partitioned into a training, validation, and testing set in order to standardize future model evaluations.

1.1 Related Work and Motivation

The release of DANSK is motivated by current limitations facing Danish NER. This introduced existing work and their shortcomings.

DaNE or Danish Named Entities (Hvingelby et al., 2020a) is an extension upon the Danish Dependency Treebank (DDT) (Nivre et al., 2016) using the CoNLL-2003 annotation standard consisting of four entity types. DaNE features high-quality annotations (inter-rater agreements of Cohen's κ=0.87 when excluding O tags) and is the dataset generally used for production ready system (Enevoldsen et al., 2021; Akbik et al., 2019; Honnibal et al., 2020).

Dan+ (Plank et al., 2021) similarly annotate DDT using the CONLL 2023 schema, but extends it further by including social media and annotating for nested named entities. With nesting, the social media domains Reddit and Twitter obtains a κ scores of 87.81 and 80.94 respectively. κ is not reported for their annotations of DDT.

Based on these sources we highlight the following limitations of Danish NER;

1. Multiple important domains such conversational speech, legal documents, web articles are currently not covered by current datasets. Moreover, even domains such as news is only covered by text spanning the period 1883-1992, thus no contemporary linguistic trends are included.

2. Current datasets are limited to the CoNLL-2003 annotation standard consisting of four entity types, as opposed to more fine-grained NER datasets like OntoNotes 5.0 which include 18 entity types, notably covered domain-specific entities such as "LAW" and does not include a "MISC", which is often excluded from evaluations (Nielsen, 2023) due to its lack of specificity.

[1] https://huggingface.co/datasets/chcaa/dansk-ner

Licensed under CC BY 4.0, pages 14–29, DOI 10.3384/nejlt.2000-1533.2024.5249

DANSK seeks to address these limitations, in part to navigate impediments to generalizability (Kirkedal et al., 2019), where domain shifts in data cause drops in performance, as models are optimized for the training and validation data, making cross-domain evaluation crucial (Plank et al., 2021). A study by Enevoldsen et al. (2021), furthermore found generalizability issues for Danish NER, not across domains, but across different types of data augmentations — further indicating generalizability issues for Danish models. Based on DANSK, we also introduce three new models of varying sizes available through DaCy 2.6.0 (Enevoldsen et al., 2021) that are specifically developed for fine-grained NER on the comprehensive array of domains included in DANSK to ensure generalizability.

Finally, we evaluate the three newly released models against some of the currently best-performing and most widely-used NLP models within Danish NER using the DANSK dataset, in order to attain estimates of generalizability across domains.

2 Dataset

2.1 The Danish Gigaword Corpus

The texts in the DANSK dataset were sampled from the Danish Gigaword Corpus (DAGW) (Strømberg-Derczynski et al., 2021), a new Danish corpus of over 1 billion words, consisting of 25 different media sources across 7 domains (see Appendix A.3.2). "Domains" within DANSK are inherited directly from the Danish Gigaword Corpus (DAGW) (Strømberg-Derczynski et al., 2021). Naturally, some domains constitute more coherent genres of text than others (e.g. "Legal" versus "Web" or "Social Media" but we have retained these labels to maintain consistency with DAGW. We take domain to refer to a distinct area or field of knowledge or activity characterized by its specific terminology, linguistic patterns, and/or unique challenges in language processing.

2.2 Initial named entity annotation

For annotation of DANSK, DAGW was filtered to exclude texts from prior to 2000 and segmented into sentences using spaCy's rule-based "sentencizer" (Honnibal et al., 2020). DANSK uses the annotation standard of OntoNotes 5.0. For NER annotation using Prodigy (Montani and Honnibal, 2018), texts were first divided up equally for the 10 annotators, with a 10% overlap between the assigned texts (i.e. 10% of texts were annotated by all annotators). The annotators were 10 native speakers of Danish (nine female, one male) between the ages of 22-30 years old, studying in the Masters degree program in English Linguistics at Aarhus University. For fine-grained NER annotation, instructions followed the 18 shorthand descriptions of the OntoNotes 5.0 named entity types (Weischedel et al., 2012). For more information on the recruitment and compensation of annotators and the annotation instruction process, see Section 8 and Appendix A.4.2.Initial annotations suffered from poor intercoder reliability, as measured by Cohen's kappa (κ) scores over tokens, calculated by matching each rater pairwise to every other (Table 1). However it has been argued that Cohen's kappa poorly reflect annotation quality due to its requirement for negative cases, and Macro F1 score has been proposed as a better alternative (Brandsen et al., 2020). The span-level Macro F1 scores were calculated for all annotators (Table 2) using the spaCy implementation (v. 3.5.4).

	Cohen's κ	
	Initial	Reviewed
Annotator 1	0.6	0.92
Annotator 2	0.52	-
Annotator 3	0.51	0.93
Annotator 4	0.58	0.93
Annotator 5	0.54	0.91
Annotator 6	0.56	0.93
Annotator 7	0.47	0.93
Annotator 8	0.51	0.89
Annotator 9	0.52	0.92
Annotator 10	0.56	-
Average		0.92

Table 1: Table showing the average Cohen's κ scores for each rater for the overlapping data after the initial annotation and after the annotations were reviewed and improved (see section 2.3).

2.3 Annotation improvement

Due to the low consensus between annotators, it was deemed necessary for the annotated texts to undergo additional processing before they could be unified into a coherent, high-quality dataset.

Texts with multiple annotators Some curated datasets utilize a single annotator for manual resolvement of conflicts between raters (Weischedel et al., 2012). While this is sometimes necessary, it skews annotations towards the opinion of a single annotator rather than the general consensus across raters. In order to resolve conflicts while diminishing this skew, we took a two-step approach: first, an automated procedure was employed to resolve the majority of annotation disagreements systematically; second, a small number of texts with remaining annotation conflicts were resolved manually.

Named-entity type	Macro F1 (Span)	SD
CARDINAL	0.47	0.23
DATE	0.55	0.21
EVENT	0.5	0.34
FACILITY	0.22	0.38
GPE	0.91	0.05
LANGUAGE	0.0	0.0
LAW	0.23	0.32
LOCATION	0.22	0.24
MONEY	0.62	0.49
NORP	0.5	0.39
ORDINAL	0.5	0.27
ORGANIZATION	0.72	0.14
PERCENT	0.0	0.0
PERSON	0.59	0.32
PRODUCT	0.12	0.23
QUANTITY	0.18	0.26
TIME	0.33	0.36
WORK OF ART	0.4	0.29

Table 2: The macro F1-scores across the raters for each of the named entity types.

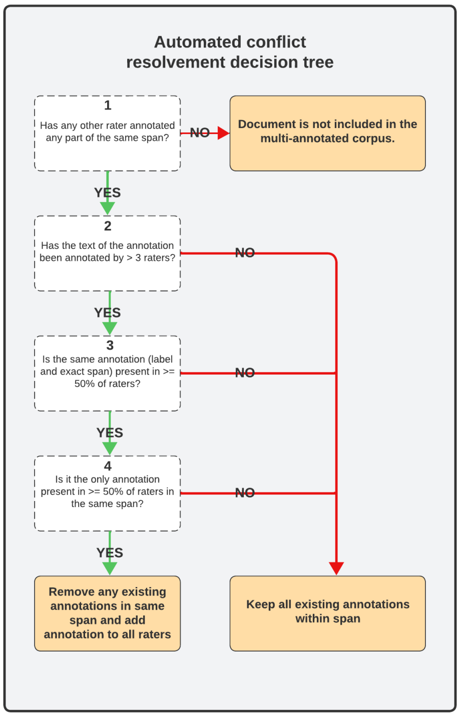
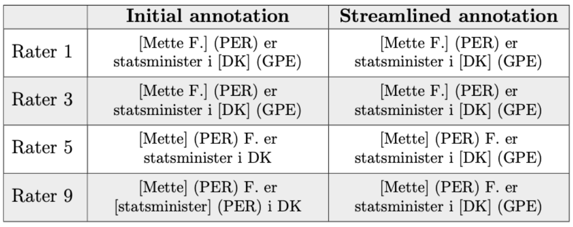

Automated conflict resolvement decision tree
1
Has any other rater annotated any part of the same span?
NO
Document is not included in the multi-annotated corpus.
YES
2
Has the text of the annotation been annotated by > 3 raters?
NO
YES
3
Is the same annotation (label and exact span) present in >= 50% of raters?
NO
YES
4
Is it the only annotation present in >= 50% of raters in the same span?
NO
YES
Remove any existing annotations in same span and add annotation to all raters
Keep all existing annotations within span

Figure 1: The decision tree for automated conflict resolvement of multi-annotated texts. Each annotation span in a text followed the steps from 1 to 4 on the diagram. The decision tree was only followed for annotation spans found in texts that had been annotated by at least four raters.



	Initial annotation	Streamlined annotation
Rater 1	[Mette F.] (PER) er statsminister i [DK] (GPE)	[Mette F.] (PER) er statsminister i [DK] (GPE)
Rater 3	[Mette F.] (PER) er statsminister i [DK] (GPE)	[Mette F.] (PER) er statsminister i [DK] (GPE)
Rater 5	[Mette] (PER) F. er statsminister i DK	[Mette] (PER) F. er statsminister i [DK] (GPE)
Rater 9	[Mette] (PER) F. er [statsminister] (PER) i DK	[Mette] (PER) F. er statsminister i [DK] (GPE)

Figure 2: An example of a text along with its four annotations being processed on the basis of the decision-tree in Figure 1.

The automated procedure for resolving annotation disagreements was rule-based and followed a decision tree-like structure (Figure 1). It was only applied to texts that had been annotated by a minimum of four raters, ensuring that that an annotation with no consensus was accepted in a text annotated by two annotators. To exemplify the streamlining of the multi-annotated texts, Figure 2 is included.

After employing the automated procedure, the 886 multi-annotated texts went from having 513 (58%) texts with complete rater agreement to 789 (89%). The texts with complete agreement were added to the DANSK dataset, while the remaining 97 (21%) of the multi-annotated texts had remaining annotation conflicts. The remaining texts with conflicting annotations were resolved manually by the first author, by changing any annotations that did not comply with the extended OntoNotes annotation guidelines. However, three texts were of such bad quality that they were rejected and excluded. The remaining resolved 94 texts were then added to DANSK.

Finally, to ensure that any named entities of the type LANGUAGE, PERCENT, and PRODUCT had not been missed by the annotators, an extra measure was taken. The model `TNER/Roberta-Large-OntoNotes5`[2] was used to add these types of annotations to the accepted multi-annotated texts (Ushio and Camacho-Collados, 2021). Each text with any predictions by the models was then manually assessed by the first author, to inspect whether the additional model annotations should be included. None of the predictions matched the annotation guidelines and were thus not added to the texts. This step concluded the processing of the multi-annotated texts, which resulted in a total of 883 texts added to the DANSK dataset.

Texts with a single annotator Based on the low consensus between the multiple raters, it was assumed that documents annotated by a single annotator might

[2] https://huggingface.co/tner/roberta-large-ontonotes5

not meet a sufficient quality standard. To refine these annotations, we utilize the reviewed annotations from multiple annotators to train a model. This model is then applied to the data such that detected discrepancies between model and human annotations are reviewed and manually resolved by the authors. The rationale for this process is that it propagates the aggregated annotations across the dataset and can thus be seen as a supervised approach to anomaly detection

As the preliminary DANSK dataset included relatively few annotations, we explored the effect of enriching our existing datasets using the English subsection of OntoNotes 5.0 (Recchia and Jones, 2009). We trained a total of three NER models using a multilingual `xlm-roberta-large`[3] to allow for cross-lingual transfer (Conneau et al., 2020): 1) the first model on 80% of the preliminary DANSK dataset; 2) the second building on (1) by adding English OntoNotes 5.0 and 3) the third duplicating the 80% of the preliminary DANSK to match the size of the English OntoNotes 5.0. All three models were validated on the remaining 20%. The best model (the third, (3)) was then applied to the remaining 15062 texts and discrepancies were manually resolved by the second author. The best model obtained an span macro-F1 of 0.80 and were trained using spaCy's transition-based parser (v2) with a batch size of 128, a gradient accumulation of 3 and a max learning rate of 5e-5 trained for 20 000 steps with 250 steps of warm-up. The remainder of the parameters were set to the default (in spaCy v. 3.5.4).

Resolving remaining inconsistencies Because of the large number of annotation reviews, we were able to identify common annotation mistakes. To further enhance the quality of the annotations, all texts were screened for common errors using a list of regex patterns (see and Appendix A.5.1). This resulted in flagged matches in 449 texts which were re-annotated in accordance with the OntoNotes 5.0 extended annotation guidelines (Weischedel et al., 2012) and the newly developed Danish Addendum designed to clarify ambiguities and issues specific to Danish texts, as described in the full dataset card (Appendix A).

3 Final dataset: DANSK

3.1 DANSK quality assessment

Average Cohen's κ scores were calculated on the processed, finalized versions of texts with multiple annotators. All of the non-removed raters' texts were included, as well as the preliminary version of DANSK with the conflicts resolved.

As expected, the average scores of the processed texts saw a marked increase, ultimately ranging between 0.93 and 0.89, compared with scores of the original annotated texts which ranged from 0.47 to 0.60 (Table 1).

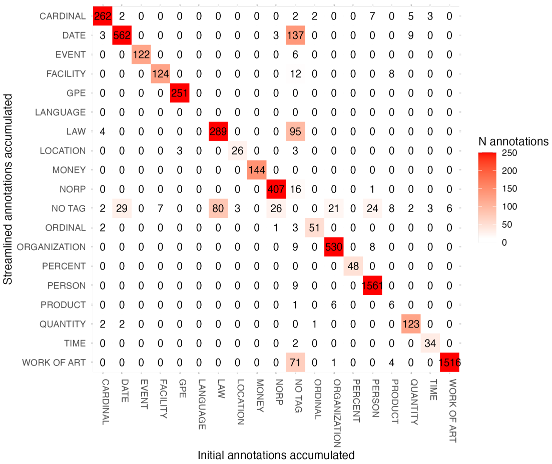


Figure 3: Confusion matrix across annotated tokens before and after the automated streamlining.

To assess which inconsistencies still remained between the DANSK dataset and the raters' annotations, a confusion matrix between the annotations of DANSK and the accumulated annotations of the processed rater texts was assessed. As can be seen in Figure 3, the majority of differences are cases in which a token or a span of tokens was considered a named entity by one party, but not by the other. In other words, no unequivocal systematic patterns between a pair of named entities existed.

To examine the final quality of the annotation process we lastly had the first author (Native speaker of Danish, Male, 29 Years) independently annotate 100 documents sampled from DANSK. These documents were sampled equally among the annotators on the non-overlapping datasets. The new annotations obtained an Span Macro-F1 of 96.6. These agreements mainly stemmed from cases which were either unclear due to too little context such as when the text was very short or cases where the labels is underspecified e.g. when a website URL (e.g. "Jobindex.dk") should be annotated as a organization.

3.2 DANSK descriptive statistics

To provide complete transparency about the dataset distributions, descriptive statistics are reported in the

[3] https://huggingface.co/xlm-roberta-large

dataset card[4] and Appendix A with regard to source, domain, and named entities. In total DANSK consists of 15 062 documents and 14 462 entities.

4 DaCy model curation

4.1 Model Specifications

In order to contribute to Danish NLP with both fine-grained tagging as well as non-domain specific performance, three new models were fine-tuned to the newly developed DANSK dataset. The three models differed in size and included a large, medium, and small model as they were fine-tuned versions of `dfm-encoder-large-v1`[5], `DanskBERT`[6] and `electra-small-nordic`[7] (Snæbjarnarson et al., 2023). These models contain 355, 278, and 22 million trainable parameters, respectively. They were chosen based on their ranking among the best-performing Danish language models within their size class, according to the ScandEval benchmark scores current as of the 7th of March, 2023 (Nielsen, 2023).

The models were all fine-tuned on the training partition of the DANSK dataset using the Python package *spaCy 3.5.0* (Honnibal et al., 2020). The fine-tuning was performed on an NVIDIA T4 GPU through the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark. An exhaustive list of all configurations of the system, as well as hyperparameter settings, is provided in the GitHub repository [8].

The three models shared the same hyperparameter settings for the training with the exception that the large model utilized an accumulated gradient of 3. They employed a batch size of 2048 and applied Adam as the optimizer with $\beta 1 = 0.9$ and $\beta 2 = 0.999$ and an initial learning rate of 0.0005. It used L2 normalization with weighted decay, $\alpha = 0.01$, and gradient clipping with c-parameter = 1.0. For the NER head of the transformer, transition-based parser (Goldberg and Nivre, 2013) was used with a hidden width of 64. The models were trained for 20,000 steps with an early stopping patience of 1600. During training the model had a dropout rate of 0.1 and an initial learning rate of 0.0005.

For the progression of the training loss of the NER head, loss of the transformer, NER performance measured in recall, precision, and F1-score, refer to the dataset card and Appendix B.

Fine-grained NER Models			
	Large	**Medium**	**Small**
F1-score	**0.823**	*0.806*	0.776
Recall	**0.834**	*0.818*	0.77
Precision	**0.813**	*0.794*	0.781

Table 3: Model performances in macro F1-scores. Bold and italics are used to represent the best and second-best scores, respectively.

Fine-grained NER Models			
Named-entity type	**Large**	**Medium**	**Small**
CARDINAL	*0.87*	0.78	**0.89**
DATE	0.85	*0.86*	**0.87**
EVENT	**0.61**	*0.57*	0.4
FACILITY	**0.55**	*0.53*	0.47
GPE	**0.89**	*0.84*	0.80
LANGUAGE	**0.90**	*0.49*	0.19
LAW	**0.69**	*0.63*	0.61
LOCATION	*0.63*	**0.74**	0.58
MONEY	*0.99*	**1**	0.94
NORP	0.78	**0.89**	*0.79*
ORDINAL	0.70	*0.7*	**0.73**
ORGANIZATION	**0.86**	*0.85*	0.78
PERCENT	*0.92*	**0.96**	**0.96**
PERSON	*0.87*	**0.87**	0.83
PRODUCT	**0.67**	*0.64*	0.53
QUANTITY	0.39	*0.65*	**0.71**
TIME	*0.64*	0.57	**0.71**
WORK OF ART	*0.49*	**0.64**	0.49
AVERAGE	**0.82**	*0.81*	0.78

Table 4: Model performances in Macro F1-scores within each named entity type. Bold and italics are used to represent the best and second-best scores, respectively.

4.2 Results

This section presents the results of the performance evaluation. An overview of the general performance of the three fine-grained models is reported in Table 3. Domain-level performance can be seen in Table 5. To account for the differences in domain size, Figure 4 is further included as it adds an additional dimension of information through the depiction of the size of the domains. Insights into performance within named entity categories are provided in Table 4.

Refer to the dataset card and Appendix A for full information on the distributions for named entities and domains within the partitions.

[4] https://huggingface.co/datasets/chcaa/dansk-ner
[5] https://huggingface.co/chcaa/dfm-encoder-large-v1
[6] https://huggingface.co/vesteinn/DanskBERT
[7] https://huggingface.co/jonfd/electra-small-nordic
[8] https://huggingface.co/datasets/chcaa/dansk-ner

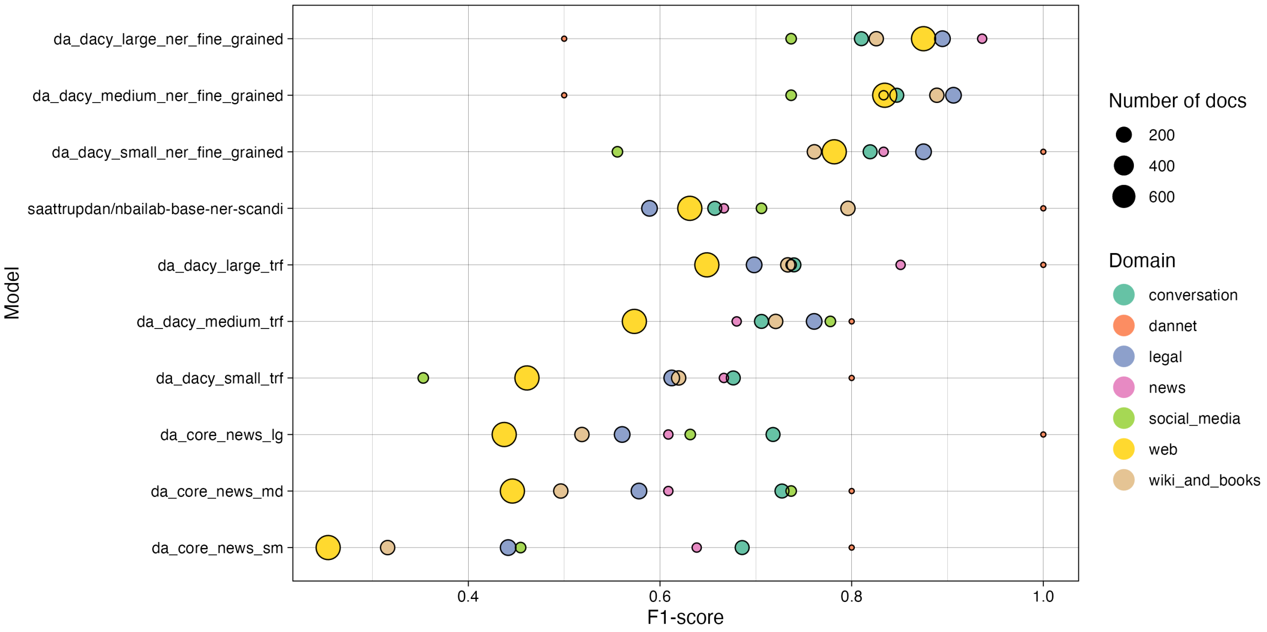


Figure 4: Domain performance in macro F1-scores of the three models on the test partition of DANSK. The size of the circles represents the size of the domains, and thus their relative weighted impact on the overall scores. See Table 5 for scores.

Fine-grained Ner Models			
Domain	**Large**	**Medium**	**Small**
All domains	**0.82**	*0.81*	0.78
Conversation	*0.80*	0.72	**0.82**
Dannet	*0.75*	0.667	**1**
Legal	*0.85*	*0.85*	**0.87**
News	*0.84*	0.76	**0.86**
Social Media	0.79	**0.85**	*0.8*
Web	**0.83**	*0.80*	0.76
Wiki and Books	*0.78*	**0.84**	0.71

Table 5: Model performances in macro F1-scores within each domain. Bold and italics are used to represent the best and second-best scores, respectively.

5 Model generalizability

5.1 Methods

5.1.1 Models

To assess whether there exists a generalizability issue for Danish language models, a number of SOTA models were chosen for evaluation on the test partition of the newly developed DANSK dataset. The field of Danish NLP and NER is evolving rapidly, making it hard to establish an overview of the most important models for Danish NER. However, the models for the evaluation were chosen on the basis of two factors; namely prominence of use, and performance. The latter was gauged on the basis of ScandEval, the NLU framework for benchmarking (Nielsen, 2023).

At the time of the model search, the model `saattrupdan/nbailab-base-ner-scandi` [9] ranked amongst the best-performing models for Danish (and Scandinavian) NER.[10] It was trained on the combined dataset of DaNE, NorNE, SUC 3.0, and the Icelandic and Faroese part of the WikiANN (Hvingelby et al., 2020b; Gustafson-Capková and Hartmann, 2006; Ejerhed et al., 1992; Jørgensen et al., 2019; Pan et al., 2017). Because of the wide palette of different datasets, texts from more domains are represented. It was thus conjectured that the model might not suffer from the generalizability issues outlined in the introduction section of the paper.

Apart from this model, the three v0.1.0 DaCy models large, medium, and small were also included. Note that these are the existing non-fine-grained models that were already in DaCy prior to the development of the fine-grained models presented in this paper. The models are fine-tuned versions of 1) Danish Ælæctra[11], Danish BERT[12], and the XLM-R (Conneau et al., 2020). The models are fine-tuned on DaNE (Hvingelby et al., 2020b) and DDT (Johannsen et al., 2015) for multitask prediction for multiple task including named-entity recognition and at the time of publication achieved state-of-the-art performance for Danish NER (Enevold-

[9] https://huggingface.co/saattrupdan/nbailab-base-ner-scandi

[10] https://paperswithcode.com/sota/named-entity-recognition-on-dane

[11] https://huggingface.co/Maltehb/aelaectra-danish-electra-small-cased

[12] https://huggingface.co/Maltehb/danish-bert-botxo

sen et al., 2021).

We also include the NLP framework *spaCy* (Explosion AI, Berlin, Germany), to explore the generalization of production systems. SpaCy features three Danish models (small, medium, and large[13]) which similarly to the DaCy models are multi-task models with NER capabilities. Although spaCy also includes a Danish transformer model, it was not incorporated in the generalizability analysis. The reason for this is that DaCy medium v.0.1.0 is already included and the two models are almost identical. Both are based on the model `Maltehb/danish-bert-botxo`[14] and fine-tuned on DaNE, and thus only deviate on minor differences in hyperparameter settings.

In summary, the models included in the final evaluation were:

1. `Base-ner-scandi` (`nbailab-base-ner-scandi`)
2. `DaCy large` (`da_dacy_large_trf-0.2.0`)
3. `DaCy medium` (`da_dacy_medium_trf-0.2.0`)
4. `DaCy small` (`da_dacy_small_trf-0.2.0`)
5. `spaCy large` (`da_core_news_lg v. 3.5.0`)
6. `spaCy medium` (`da_core_news_md v. 3.5.0`)
7. `spaCy small` (`da_core_news_sm v. 3.5.0`)
8. `Fine-grained large` (`da_dacy_large_trf-0.1.0`)
9. `Fine-grained medium` (`da_dacy_medium_trf-0.1.0`)
10. `Fine-grained small` (`da_dacy_small_trf-0.1.0`)

5.1.2 Named Entity Label Transfer

A fine-grained NER dataset with 18 labels following the OntoNotes guidelines has not been publicly available for Danish until now. The aforementioned models have thus only been fine-tuned to the classic, more coarse-grained DaNE dataset that follows the CoNLL-2003 named entity annotation scheme (Sang and De Meulder, 2003; Hvingelby et al., 2020a). This includes the four named entity types PER (person), LOC (location), ORG (organization), and MISC (miscellaneous). This annotation scheme is radically different from the DANSK annotations that match the OntoNotes 5.0 standards. To enable an evaluation of the models, the DANSK named entity labels were coerced into the CoNLL-2003 standard in order to match the nature of the models, and specifically to assist us in highlighting performance disparities across out-of-distribution domains, such as ”SoMe” and ”Legal”, which are new in the release of DaNSK.

As the description of both ORG and PER in CoNLL-2003 largely matches that of the extended OntoNotes, these named entity types could be used in the evaluation with a 1-to-1 mapping without further handling. However, in CoNLL-2003, LOC includes cities, roads, mountains, abstract places, specific buildings, and meeting points (Hvingelby et al., 2020a; Sang and De Meulder, 2003). As the extended OntoNotes guidelines use both GPE and LOCATION, DANSK GPE annotations were mapped to LOC in an attempt to make the test more accurate. Predictions for the CoNLL-2003 MISC category, intended for names not captured by other categories (e.g. events and adjectives such as ”*2004 World Cup*” and ”*Italian*”), were excluded.

5.1.3 Evaluation

SOTA models were evaluated using macro average F1-statistics at a general level, a domain level, and finally F1-scores at the level of named entity types.

5.2 Results

Table 6 provides an overview of macro span-F1-scores as well as recall and precision statistics. The performance across domains and across named entity types are reported in Table 7 and Table 8.

Model	F1	Recall	Precision
Base-ner-scandi	*0.64*	0.59	**0.70**
DaCy large	**0.68**	**0.67**	*0.69*
DaCy medium	0.63	*0.64*	0.61
DaCy small	0.51	0.48	0.56
spaCy large	0.49	0.45	0.53
spaCy medium	0.49	0.47	0.52
spaCy small	0.32	0.32	0.32

Table 6: Overall performance on the DANSK test set in macro F1-score using the CoNLL-2003 Schema. Bold and italic represent the best and next best scores.

6 Discussion

6.1 DANSK dataset

The DANSK dataset enhances Danish NER by focusing on fine-grained named entity labels and diverse domains like conversations, legal matters, and web sources, but omits some domains in DaNE, such as magazines (Norling-Christensen, 1998; Hvingelby et al., 2020a). Entity distribution varies, influencing model performance for specific types.

DANSK's quality was benchmarked using models trained on different OntoNotes 5.0 annotated datasets (Luoma et al., 2021). Despite the dataset size disparity, performances for English and Finnish models were between F1-scores of .89 and .93 (Luoma et al., 2021; Li et al., 2022), notably higher than DANSK. Given the smaller size of DANSK (15062 texts) compared to English OntoNotes (600000 texts) (Weischedel et al.,

[13] Note that a model size of spaCy are not comparable to model sizes of transformer encoders

[14] https://huggingface.co/Maltehb/danish-bert-botxo

Model	Across	Conversational	Legal	News	SoMe	Web	Wiki
base-ner-scandi	*0.64*	0.66	0.59	*0.67*	0.71	*0.63*	**0.80**
DaCy Large	**0.68**	**0.74**	*0.70*	**0.85**	*0.74*	**0.65**	*0.73*
DaCy Medium	0.63	0.71	**0.76**	0.68	**0.78**	0.57	0.72
DaCy Small	0.51	0.68	0.61	*0.67*	0.35	0.46	0.62
spaCy Large	0.49	0.72	0.56	0.61	0.63	0.44	0.52
spaCy Medium	0.49	*0.73*	0.58	0.61	*0.74*	0.45	0.50
spaCy small	0.32	0.69	0.44	0.64	0.46	0.25	0.32

Table 7: The domain-level performances in macro F1-scores on the DANSK test set using the CoNLL-2003 Schema. Bold and italic represent the best and next best scores.

2012), performance for models trained on DANSK is expectedly lower, irrespective of annotation quality (Russakovsky et al., 2015).

Annotation quality issues were tackled, improving Cohen's κ values from ~0.5 to ~0.9 (Table 1 and Table ??). Initial difficulties arose from suboptimal sampling from DAGW and insufficient annotator training. Future improvements include initial quality screening and comprehensive training with the OntoNotes 5.0 annotation scheme (Plank, 2022; Uma et al., 2021). In the release of the DANSK dataset, we include raw (per annotator) annotations to allow for transparency and further analysis of annotator disagreement.

6.2 DaCy models

New fine-grained models of varying sizes attained macro F1-scores of 0.82, 0.81, and 0.78 respectively. Larger models generally performed better as would be expected. However, due to DANSK's domain imbalance, these scores should be treated carefully. Domains like web, conversation, and legal heavily influenced the F1-scores due to their larger text volume. Performance comparisons are based on OntoNotes 5.0 standard datasets due to the unique annotation scheme of DANSK.

Minor performance variation was found within each domain. The small models excelled in underrepresented domains like news, possibly leading to volatile results. Legal texts were easiest to classify with F1-scores of 0.85 and 0.87.

Classification performance varied with named entity types. Facilities, artworks, and quantities were difficult to predict, whereas entities like money, dates, percentages, GPEs, organizations, and cardinals were easier to classify. This can be attributed to the quantity and context of named entities in the training data. Some entity types might appear in similar contexts or have similar structures, hence easier to distinguish. Variance in performance may arise from differences in text quality and context. Given the observed performance differences across domains and named entity types, it's crucial to understand the strengths and limitations of the new models within the DaCy framework.

6.3 SOTA models and generalizability

The new fine-grained DaCy models demonstrate higher performance on the DANSK dataset, compared to existing SOTA models (refer to Tables 6 and 3). However, due to annotation scheme discrepancies, a direct comparison is challenging.

Performance analysis is two-fold: evaluation across domains for each model, and comparison between models, both following the CoNLL-2003 annotation scheme.

Significant domain performance disparities were observed (see Table 7). For instance, `base-ner-scandi` scored F1-scores of 0.59 and 0.8 for legal and Wikipedia texts, respectively. Actual model accuracy may vary by domain, contrary to performance reported on DaNE. The models performed best on conversation and news texts, with web and wiki sources performing poorly.

Larger models generally outperformed smaller models, with `base-ner-scandi` and `DaCy large` performing best, with across-domain F1-scores of 0.64 and 0.68 respectively. The DaCy models, easily accessible via the DaCy framework, performed comparably or better than the `base-ner-scandi` model, hence DaCy is the preferred library for Danish NER.

Table 8 shows the performance of models within each non-fine-grained named entity class (CoNLL-2003) on the DaNSK test set, and includes scores for the previously best-performing non-fine-grained DaCy models (0.2.0). The release of fine-grained NER DaCy models (0.1.0) represents a significant performance improvement, from an overall average F1-score of 0.67 for DaCy Large (0.2.0) versus 0.85 for DaCy fine-grained large (0.1.0).

7 Conclusion

Danish NER suffers from limited dataset availability, lack of cross-validation, domain-specific evaluations, and fine-grained NER annotations. This paper intro-

Model	Average F1	Person F1	Organization F1	Location F1
DaCy large	0.67 (0.62, 0.72)	0.74 (0.67, 0.80)	0.50 (0.43, 0.57)	0.80 (0.73, 0.86)
DaCy medium	0.56 (0.49, 0.60)	0.62 (0.54, 0.68)	0.40 (0.32, 0.47)	0.66 (0.53, 0.75)
DaCy small	0.55 (0.50, 0.59)	0.64 (0.56, 0.71)	0.38 (0.31, 0.46)	0.65 (0.56, 0.72)
base-ner-scandi	0.64 (0.57, 0.69)	0.69 (0.62, 0.77)	0.49 (0.38, 0.59)	0.72 (0.58, 0.81)
SpaCy large	0.51 (0.43, 0.56)	0.60 (0.52, 0.68)	0.33 (0.24, 0.42)	0.61 (0.46, 0.71)
SpaCy Medium	0.50 (0.44, 0.55)	0.59 (0.51, 0.65)	0.32 (0.26, 0.41)	0.62 (0.48, 0.72)
SpaCy Small	0.34 (0.30, 0.40)	0.36 (0.29, 0.43)	0.22 (0.16, 0.29)	0.46 (0.35, 0.55)
Fine-grained large (ours)	**0.85** (0.81, 0.88)	*0.86* (0.80, 0.90)	*0.79* (0.73, 0.85)	**0.93** (0.89, 0.96)
Fine-grained medium (ours)	**0.85** (0.81, 0.88)	0.85 (0.79, 0.90)	**0.80** (0.76, 0.85)	*0.91* (0.86, 0.96)
Fine-grained small (ours)	*0.83* (0.80, 0.86)	**0.87** (0.82, 0.92)	*0.79* (0.74, 0.83)	0.85 (0.78, 0.92)

Table 8: Performance using the CoNLL-2003 Schema in F1-scores on the DaNSK test set. Bold and italic represent the best and next best scores. Scores are bootstrapped on the documents level and shows the mean the 95% confidence interval in showed in the parentheses.

duces DANSK, a high-granularity named entity dataset for within-domain evaluation, DaCy 2.6.0 with three generalizable, fine-grained models, and an evaluation of contemporary Danish models. DANSK, annotated following OntoNotes 5.0 and including metadata on text origin, facilitates across-domain evaluations. However, observed performance still falls short of what is seen among higher-resourced languages. DaCy models, trained on DANSK, achieve up to 0.82 macro F1-score on fine-grained NER across 18 categories. While work remains to be done to augment the size and quality of fine-gained named entity annotation in Danish, the release of DANSK and DaCy will assist in addressing generalizability issues in the field.

8 Ethics statement

Ethics Statement Our dataset is constructed based on the public dataset The Danish Gigaword corpus, which followed ethical practices in its composition. For spoken conversations, participants agreed on releasing anonymized transcripts of their conversations. Social media data only includes publicly available Tweets. Because distribution of this part of the dataset is through Tweet IDs and requires rehydration, any Tweets subsequently removed by the user are no longer included.

10 native Danish speakers enrolled in the English Linguistics Master's program were recruited as annotators through announcements in classrooms. This degree program was chosen because students receive relevant training in general linguistics, including syntactic analysis. We employed a larger group of students to adhere to institutional limitations on the number of hours student workers can have. The demographic bias of our annotators (nine female, one male) reflects the demographics of this MA program. Annotators worked 10 hours/week for six weeks from October 11, 2021, to November 22, 2021. Their annotation tasks included part-of-speech tagging, dependency parsing, and NER annotation. Annotators were compensated at the standard rate for students, as determined by the collective agreement of the Danish Ministry of Finance and the Central Organization of Teachers and the CO10 Central Organization of 2010 (the CO10 joint agreement), which is 140DKK/hour.

We are committed to full transparency and replicability in our release of DaNSK. Following work by Mitchell et al. (2019) and (Gebru et al., 2021), we provide a dataset card for DANSK following the format proposed in Lhoest et al. (2021), which can be accessed here: https://huggingface.co/datasets/chcaa/dansk-ner. The dataset card and additional supporting information about the language resource will also be included in the Appendices upon publication.

Acknowledgements

All of the computation done for this project was performed on the UCloud interactive HPC system, which is managed by the eScience Center at the University of Southern Denmark.

References

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Brandsen, Alex, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. 1992. The linguistic annotation system of the stockholm-umeå project. *Department of Linguistics, University of Umeå, Sweden*.

Enevoldsen, Kenneth, Lasse Hansen, and Kristoffer Nielbo. 2021. Dacy: A unified framework for danish nlp. *arXiv preprint arXiv:2107.05295*.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Goldberg, Yoav and Joakim Nivre. 2013. Training deterministic parsers with non-deterministic oracles. *Transactions of the association for Computational Linguistics*, 1:403–414.

Gustafson-Capková, Sofia and Britt Hartmann. 2006. Manual of the stockholm umeå corpus version 2.0. *Unpublished Work*, pages 5–85.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python. Publisher: Zenodo, Honolulu, HI, USA.

Hvingelby, Rasmus, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020a. Dane: A named entity resource for danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4597–4604.

Hvingelby, Rasmus, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020b. Dane: A named entity resource for danish. In *Proceedings of the 12th language resources and evaluation conference*, pages 4597–4604.

Johannsen, Anders, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 157.

Jørgensen, Fredrik, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2019. NorNE: Annotating named entities for norwegian. *arXiv preprint arXiv:1911.12146*.

Kirkedal, Andreas, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The lacunae of danish natural language processing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 356–362.

Lhoest, Quentin, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. 2021. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.

Li, Xiaoya, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2022. A unified MRC framework for named entity recognition.

Luoma, Jouni, Li-Hsin Chang, Filip Ginter, and Sampo Pyysalo. 2021. Fine-grained named entity annotation for finnish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 135–144.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Montani, Ines and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence to appear*.

Nielsen, Dan Saattrup. 2023. Scandeval: A benchmark for scandinavian natural language processing. *arXiv preprint arXiv:2304.00906*.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666. European Language Resources Association (ELRA).

Norling-Christensen, Ole & Asmussen Jorg. 1998. The corpus of the danish dictionary. *Lexikos*, 8(8):223–242. Publisher: Bureau of the WAT.

Pan, Xiaoman, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Plank, Barbara. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation.

Plank, Barbara, Kristian Nørgaard Jensen, and Rob van der Goot. 2021. DaN+: Danish nested named entities and lexical normalization.

Recchia, Gabriel and Michael N. Jones. 2009. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, 41:647–656. Publisher: Springer.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252. Publisher: Springer.

Sang, Erik F. and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Snæbjarnarson, Vésteinn, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Tórshavn, Faroe Islands. Linköping University Electronic Press, Sweden.

Strømberg-Derczynski, Leon, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The danish gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421. Linköping University Electronic Press, Sweden.

Uma, Alexandra, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347. Association for Computational Linguistics.

Ushio, Asahi and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62. Association for Computational Linguistics.

Weischedel, Ralph, Pradeer Sameer, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, and Mohammed El-Bachouti. 2012. OntoNotes release 5.0. *LDC Catalog*.

A Dataset card

Following work by Mitchell et al. (2019) and (Gebru et al., 2021), we provide a dataset card for DANSK following the format proposed in Lhoest et al. (2021), which can be accessed here: https://huggingface.co/datasets/chcaa/dansk-ner

A.1 Dataset Summary

DANSK: Danish Annotations for NLP Specific TasKs is a dataset consisting of texts from multiple domains, sampled from the Danish GigaWord Corpus (DAGW).[15] The dataset was created to fill in the gap of Danish NLP datasets from different domains, that are required for training models that generalize across domains. The Named-Entity annotations are moreover fine-grained and have a similar form to that of OntoNotes v5, which significantly broadens the use cases of the dataset. The domains include Web, News, Wiki & Books, Legal, Dannet, Conversation and Social Media. For a more in-depth understanding of the domains, please refer to DAGW.

The distribution of texts and Named Entities within each domain can be seen in the table below:

A.1.1 Update log

- 2023-05-26: Added individual annotations for each annotator to allow for analysis of inter-annotator agreement

A.1.2 Supported Tasks

The DANSK dataset currently only supports Named-Entity Recognition, but additional version releases will contain data for more tasks.

A.1.3 Languages

All texts in the dataset are in Danish. Slang from various platforms or dialects may appear, consistent with the domains from which the texts originally have been sampled - e.g. Social Media.

[15] Note that DAGW is not part of the Linguistic Data Consortium family of Gigaword corpora, and has notable differences in its source and composition.

A.2 Dataset Structure

A.2.1 Data Instances

The JSON-formatted data is in the form seen below:

```
{
    "text": "Aborrer over 2 kg er en uhyre sj\u00e6lden fangst.",
    "ents": [{"start": 13, "end": 17, "label": "QUANTITY"}],
    "sents": [{"start": 0, "end": 45}],
    "tokens": [
        {"id": 0, "start": 0, "end": 7},
        {"id": 1, "start": 8, "end": 12},
        {"id": 2, "start": 13, "end": 14},
        {"id": 3, "start": 15, "end": 17},
        {"id": 4, "start": 18, "end": 20},
        {"id": 5, "start": 21, "end": 23},
        {"id": 6, "start": 24, "end": 29},
        {"id": 7, "start": 30, "end": 37},
        {"id": 8, "start": 38, "end": 44},
        {"id": 9, "start": 44, "end": 45},
    ],
    "spans": {"incorrect_spans": []},
    "dagw_source": "wiki",
    "dagw_domain": "Wiki & Books",
    "dagw_source_full": "Wikipedia",
}
```

A.2.2 Data Fields

- `text`: The text
- `ents`: The annotated entities
- `sents`: The sentences of the text
- `dagw_source`: Shorthand name of the source from which the text has been sampled in the Danish Gigaword Corpus
- `dagw_source_full`: Full name of the source from which the text has been sampled in the Danish Gigaword Corpus
- `dagw_domain`: Name of the domain to which the source adheres to

A.2.3 Data Splits

The data was randomly split up into three distinct partitions; train, dev, as well as a test partition. The splits come from the same pool, and there are thus no underlying differences between the sets. To see the distribution of named entities, and domains of the different partitions, please refer to the paper, or read the superficial statistics provided in the Dataset composition section.

A.3 Descriptive Statistics

A.3.1 Dataset Composition

Named entity annotation composition across partitions is provided in Table 9.

A.3.2 Domain distribution

"Domains" within DANSK are inherited directly from the Danish Gigaword Corpus (DAGW) (Strømberg-Derczynski et al., 2021). Naturally, some domains constitute more coherent genres of text than others (e.g. "Legal" versus "Web" or "Social Media" but we have retained these labels to maintain consistency with DAGW. We take domain to refer to a distinct area or field of knowledge or activity characterized by its specific terminology, linguistic patterns, and/or unique challenges in language processing.

Domain and source distribution across partitions is provided in Table 10.

A.3.3 Entity Distribution across partitions

Domain and named entity distributions for the training, testing, and validation sets can be found in the full dataset card accompanying DANSK: https://huggingface.co/datasets/chcaa/dansk-ner

A.4 Dataset Creation

A.4.1 Curation Rationale

The dataset is meant to fill in the gap of Danish NLP that up until now has been missing a dataset with 1) fine-grained named entity recognition labels, and 2) high variance in domain origin of texts. As such, it is the intention that DANSK should be employed in training by anyone who wishes to create models for NER that are both generalizable across domains and fine-grained in their predictions. It may also be utilized to assess across-domain evaluations, in order to unfold any potential domain biases. While the dataset currently only entails annotations for named entities, it is the intention that future versions of the dataset will feature dependency Parsing, pos tagging, and possibly revised NER annotations.

A.4.2 Annotations

Annotation process To afford high granularity, the DANSK dataset utilized the annotation standard of OntoNotes 5.0, featuring 18 different named entity types. The full description can be seen in the associated paper.

Annotators 10 native Danish speakers enrolled in the English Linguistics Master's program were recruited through announcements in classrooms. This degree program was chosen because students receive relevant training in general linguistics, including syntactic analysis. We employed a larger group of students to adhere to institutional limitations on the number of hours student workers can have. The demographic bias of our annotators (nine female, one male) reflects the demographics of this MA program. Annotators worked 10 hours/week for six weeks from October 11, 2021, to November 22, 2021. Their annotation tasks included part-of-speech tagging, dependency parsing, and NER

Table 9: Named entity annotation composition across partitions

	Full	**Train**	**Validation**	**Test**
Texts	15062	12062 (80%)	1500 (10%)	1500 (10%)
Named entities	14462	11638 (80.47%)	1327 (9.18%)	1497 (10.25%)
CARDINAL	2069	1702 (82.26%)	168 (8.12%)	226 (10.92%)
DATE	1756	1411 (80.35%)	182 (10.36%)	163 (9.28%)
EVENT	211	175 (82.94%)	19 (9.00%)	17 (8.06%)
FACILITY	246	200 (81.30%)	25 (10.16%)	21 (8.54%)
GPE	1604	1276 (79.55%)	135 (8.42%)	193 (12.03%)
LANGUAGE	126	53 (42.06%)	17 (13.49%)	56 (44.44%)
LAW	183	148 (80.87%)	17 (9.29%)	18 (9.84%)
LOCATION	424	351 (82.78%)	46 (10.85%)	27 (6.37%)
MONEY	714	566 (79.27%)	72 (10.08%)	76 (10.64%)
NORP	495	405 (81.82%)	41 (8.28%)	49 (9.90%)
ORDINAL	127	105 (82.68%)	11 (8.66%)	11 (8.66%)
ORGANIZATION	2507	1960 (78.18%)	249 (9.93%)	298 (11.87%)
PERCENT	148	123 (83.11%)	13 (8.78%)	12 (8.11%)
PERSON	2133	1767 (82.84%)	191 (8.95%)	175 (8.20%)
PRODUCT	763	634 (83.09%)	57 (7.47%)	72 (9.44%)
QUANTITY	292	242 (82.88%)	28 (9.59%)	22 (7.53%)
TIME	218	185 (84.86%)	18 (8.26%)	15 (6.88%)
WORK OF ART	419	335 (79.95%)	38 (9.07%)	46 (10.98%)

annotation. Annotators were compensated at the standard rate for students, as determined by the collective agreement of the Danish Ministry of Finance and the Central Organization of Teachers and the CO10 Central Organization of 2010 (the CO10 joint agreement), which is 140DKK/hour. Named entity annotations and dependency parsing was done from scratch, while the POS tagging consisted of corrections of silver-standard predictions by an NLP model.

A.5 Automatic correction

During the manual correction of the annotation a series of consistent errors were found. These were corrected using Regex patterns (in Appendix A.5.1) which can also be viewed in full with the DANSK release along with the Danish Addendum to the Ontonotes annotation guidelines: https://huggingface.co/datasets/chcaa/dansk-ner.

.

Table 10: Domain and source distribution across partitions

Domain	Source	Full	Train	Dev	Test
Conversation	Europa Parlamentet	206	173	17	16
Conversation	Folketinget	23	21	1	1
Conversation	NAAT	554	431	50	73
Conversation	OpenSubtitles	377	300	39	38
Conversation	Spontaneous speech	489	395	54	40
Dannet	Dannet	25	18	4	3
Legal	Retsinformation.dk	965	747	105	113
Legal	Skat.dk	471	364	53	54
Legal	Retspraktis	727	579	76	72
News	DanAvis	283	236	20	27
News	TV2R	138	110	16	12
Social Media	hestenettet.dk	554	439	51	64
Web	Common Crawl	8270	6661	826	783
Wiki & Books	adl	640	517	57	66
Wiki & Books	Wikipedia	279	208	30	41
Wiki & Books	WikiBooks	335	265	36	34
Wiki & Books	WikiSource	455	371	43	41

A.5.1 Regex patterns

```
For matching with TIME spans, e.g. [16:30 - 17:30] (TIME):
\d{1,2}:\d\d ?[-|\||\/] ?\d
dag: \d{1,2}

For matching with DATE spans, e.g. [1938 - 1992] (DATE):
\d{2,4} ?[-|{] ?\d{2,4}

For matching companies with A/S og ApS,
e.g. [Hansens Skomager A/S] (ORGANIZATION):
ApS
A\/S

For matching written numerals, e.g. "en":
to | to$|^to| To | To$|^To| TO | TO$|^TO|
tre | tre$|^tre| Tre | Tre$|^Tre| TRE | TRE$|^TRE|
fire | fire$|^fire| Fire | Fire$|^Fire| FIRE | FIRE$|^FIRE|
fem | fem$|^fem| Fem | Fem$|^Fem| FEM | FEM$|^FEM|
seks | seks$|^seks| Seks | Seks$|^Seks| SEKS | SEKS$|
^SYV|
otte | otte$|^otte| Otte | Otte$|^Otte| OTTE | OTTE$|^OTTE|
ni | ni$|^ni| Ni | Ni$|^Ni| NI | NI$|^NI|
ti | ti$|^ti| Ti | Ti$|^Ti| TI | TI$|^TI

For matching "Himlen" or "Himmelen" already annotated
as LOCATION, e.g. "HIMLEN":
[Hh][iI][mM][lL][Ee][Nn]|[Hh][iI][mM][mM][Ee][lL][Ee][Nn]

For matching "Gud" already tagged as PERSON, e.g. "GUD":
[Gg][Uu][Dd]

For matching telephone numbers wrongly already
tagged as CARDINAL, e.g. "20 40 44 30":
\d{2} \d{2} \d{2} \d{2}
\+\d{2} \d{2} ?\d{2} ?\d{2} ?\d{2}$
\+\d{2} \d{2} ?\d{2} ?\d{2} ?\d{2}$
 \d{4} ?\d{4}$
^\d{4} ?\d{4}$

For matching websites already
wrongly tagged as ORGANIZATION:
.dk$|.com$

For matching Hotels and Resorts
already wrongly tagged as ORGANIZATION:
.*[h|H]otel.*|.*[R|r]esort.*

For matching numbers including /
or :, already wrongly tagged as CARDINAL:
\/
\/
-
For matching rights already
wrongly tagged as LAW:
[C|c]opyright
[®|©]
[f|F]ortrydelsesret
[o|O]phavsret$
enneskeret
```

A.6 Licensing Information

Creative Commons Attribution-Share Alike 4.0 International license

B Training progression

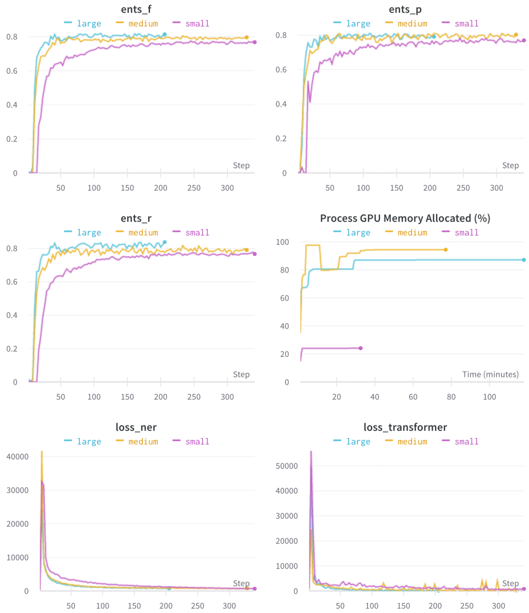


Figure 5: The epoch training progression of loss of the NER head (loss_ner), loss of the transformer (loss_transformer), NER performance measured in recall (ents_r), precision (ents_p), F1-score (ents_f) and GPU-allocation percentage.

Understanding Counterspeech for Online Harm Mitigation

Yi-Ling Chung,[*] The Alan Turing Institute, UK yilingchung27@gmail.com
Gavin Abercrombie, The Interaction Lab, Heriot-Watt University, UK g.abercrombie@hw.ac.uk
Florence Enock, The Alan Turing Institute, UK fenock@turing.ac.uk
Jonathan Bright, The Alan Turing Institute, UK bright@turing.ac.uk
Verena Rieser,[†] The Interaction Lab, Heriot-Watt University, UK v.t.rieser@hw.ac.uk

Abstract Counterspeech offers direct rebuttals to hateful speech by challenging perpetrators of hate and showing support to targets of abuse. It provides a promising alternative to more contentious measures, such as content moderation and deplatforming, by contributing a greater amount of positive online speech rather than attempting to mitigate harmful content through removal. Advances in the development of large language models mean that the process of producing counterspeech could be made more efficient by automating its generation, which would enable large-scale online campaigns. However, we currently lack a systematic understanding of several important factors relating to the efficacy of counterspeech for hate mitigation, such as which types of counterspeech are most effective, what are the optimal conditions for implementation, and which specific effects of hate it can best ameliorate. This paper aims to fill this gap by systematically reviewing counterspeech research in the social sciences and comparing methodologies and findings with natural language processing (NLP) and computer science efforts in automatic counterspeech generation. By taking this multi-disciplinary view, we identify promising future directions in both fields.

1 Introduction

The exposure of social media users to online hate and abuse continues to be a cause for public concern. Volumes of abuse on social media continue to be significant in absolute terms (Vidgen et al., 2019), and some claim they are rising on platforms such as Twitter where, at the same time, content moderation appears to be becoming less of a priority (Frenkel and Conger, 2022). Receiving abuse can have negative effects on the mental health of targets, and also on others witnessing it (Siegel, 2020; Saha et al., 2019). In the context of public figures, the impact on the witnesses (bystanders) is arguably even more important, as the abuse is potentially witnessed by a large volume of people. In addition, politicians and other prominent actors are driven out of the public sphere precisely because of the vitriol they receive on a daily basis (News, 2018), raising concerns for the overall health of democracy.

Within this context, research on mechanisms for combating online abuse is becoming ever more important. One such research angle is the area of "counterspeech" (or counter-narratives): content that is designed to resist or contradict abusive or hateful content (Benesch, 2014a; Saltman and Russell, 2014; Bartlett and Krasodomski-Jones, 2015), also see Figure 1. Such coun-

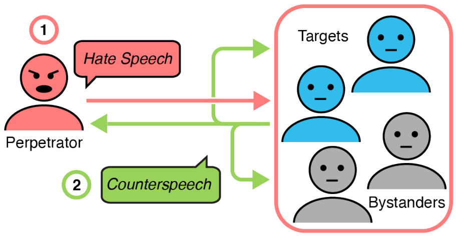


Figure 1: Counterspeech dynamics. (1) Perpetrator(s) generate Hate Speech. This may be witnessed by either targets and/or bystanders. (2) Counterspeaker(s) respond with counterspeech, which may be directed at the perpetrator(s), bystanders (e.g. to provide alternative perspectives), or other targets (e.g. in support). Counterspeakers may themselves be targets or bystanders, or could be members of organised counterspeech groups. They can have *in*- or *out*-group identities with respect to either the perpetrator(s) or the target(s). Counterspeech is directed at recipients, who can be one or more of (a) the perpetrator(s), (b) the target(s), or (c) other bystanders. Both counterspeakers and targets can be individual or multiple (one-to-one, one-to-many and so on).

[*] Now at Genaios Safe AI.
[†] Now at Google DeepMind.

Licensed under CC BY 4.0, pages 30–49, DOI 10.3384/nejlt.2000-1533.2024.5203

terspeech (as we will elaborate more fully below) is an important potential tool in the fight against online hate and abuse as it does not require any interventions from the platform or from law enforcement, and may contribute to mitigating the effects of abuse (Munger, 2017; Buerger, 2021b; Hangartner et al., 2021; Bilewicz et al., 2021) without impinging on free speech. Several civil organisations have used counterspeech to directly challenge hate, and Facebook has launched campaigns with local communities and policymakers to promote accessibility to counterspeech tools.[1] Similarly, Moonshot and Jigsaw implemented The Redirect Method, presenting alternative counterspeech or counter videos when users search queries that may suggest an inclination towards extremist content or groups.[2]

The detection and generation of counterspeech is important because it underpins the promise of AI-powered assistive tools for hate mitigation. Identifying counterspeech is vital also for analytical research in the area: for instance, to disentangle the dynamics of perpetrators, victims and bystanders (Mathew et al., 2018; Garland et al., 2020, 2022), as well as determining which responses are most effective in combating hate speech (Mathew et al., 2018, 2019; Chung et al., 2021a).

Automatically producing counterspeech is a timely and important task for two reasons. First, composing counterspeech is time-consuming and requires considerable expertise to be effective (Chung et al., 2021b). Recently, large language models have been able to produce fluent and personalised arguments tailored to user expectations addressing various topics and tasks. Thus, developing counterspeech tools is feasible and can provide support to civil organisations, practitioners and stakeholders in hate intervention at scale. Second, by partially automating counterspeech writing, such assistive tools can lessen practitioners' psychological strain resulting from prolonged exposure to harmful content (Riedl et al., 2020; Chung et al., 2021b).

However, despite the potential for counterspeech, and the growing body of work in this area, the research agenda remains a relatively new one, which also suffers from the fact that it is divided into a number of disciplinary silos. In methodological terms, meanwhile, social scientists studying the dynamics and impacts of counterspeech (e.g. Munger, 2017; Buerger, 2021b; Hangartner et al., 2021; Bilewicz et al., 2021) often do not engage with computer scientists developing models to detect and generate such speech (e.g. Chung et al., 2021c; Saha et al., 2022) (or vice versa). This disconnection may increase the time and effort for tackling online harms.

The aim of this review article is to fill this gap, by providing a comprehensive, multi-disciplinary overview of the field of counterspeech covering computer science[3] and the social sciences over the last ten years. We make a number of contributions in particular. Firstly, we outline a definition of counterspeech and a framework for understanding its use and impact, as well as a detailed taxonomy. Visualised in Figure 1, such a framework helps delineate the interaction of hate speech and responses within people involved in the conversations (i.e. perpetrators, targets and bystanders). We review research on the effectiveness of counterspeech, bringing together perspectives on the impact it makes when it is experienced. Thus, computer scientists can adeptly approach counterspeech studies and develop effective tools based on our analysis. We also analyse technical work on counterspeech, looking specifically at the task of counterspeech generation, scalability, and the availability and methodology behind different datasets. Importantly, across all studies, we focus on commonalities and differences between computer science and the social sciences, including how the impact of counterspeech is evaluated and which specific effect of hate speech it best ameliorates.

We draw on our findings to discuss the challenges and directions of open science (and safe AI) for online hate mitigation. For computer scientists, we provide evidence-based recommendations for automatic approaches to counterspeech tools using Natural Language Processing (NLP). Similarly, for social scientists, we set out future perspectives on interdisciplinary collaborations with AI researchers on mitigating online harms, including conducting large-scale analyses and evaluating the impact of automated interventions. Taken together, our work offers researchers, policymakers and practitioners the tools to further understand the potentials of automated counterspeech for online hate mitigation.

2 Background

Interest in investigating the social and computational aspects of counterspeech has grown considerably in the past five years. However, while extant work reviews the impact of counterspeech on hate mitigation (Saltman and Russell, 2014; Carthy et al., 2020; Buerger, 2021a), none have systematically addressed this issue in combination with computational studies in order to synthesise social scientific insights and discuss the potential role of automated methods in reducing harms. Carthy et al. (2020) present a focused (2016-2018) systematic review of research into the impact of counter-narratives on prevention of violent radicalisation. They cate-

[1] https://counterspeech.fb.com/en/

[2] https://moonshotteam.com/the-redirect-method/

[3] While most studies on computational approaches to counterspeech included in this review adopt natural language processing techniques, we use 'computer science' to broadly cover the research field in which the studies are done.

gorise the techniques employed in counter-narratives into four groups: (1) counter-stereotypical exemplars (challenging stereotypes, social schema or moral exemplars), (2) persuasion (e.g., through role-playing and emotion inducement), (3) inoculation (proactively reinforcing resistance to attitude change or persuasion), and (4) alternative accounts (disrupting false beliefs by offering different perspectives of events). The measurements of counter-narrative interventions are based on (1) intent of violent behaviour, (2) perceived symbolic/realistic group threat (e.g., perception of an outgroup as dangerous), and (3) in-group favouritism/outgroup hostility (e.g., level of trust, confidence, discomfort and forgiveness towards out-groups). They argue that counter-narratives show promise in reducing violent radicalisation, while its effects vary across techniques, with counter-stereotypical exemplars, inoculation and alternative accounts demonstrating the most noticeable outcomes. Buerger (2021a) reviews the research into the effectiveness of counterspeech, attempting to categorise different forms of counterspeech, summarise the source of influences in abusive/positive behaviour change, and elucidate the reasons which drive strangers to intervene in cyberbullying. Here, the impact of counterspeech is mostly evaluated by the people involved in hateful discussions, including hateful speakers, audiences, and counterspeakers. In comparison, we focus on *what* makes counterspeech effective by comprehensively examining its use based on aspects such as strategies, audience and evaluation.

On the computational side, some work reviews the use of counterspeech in social media using natural language processing, including work outlining counterspeech datasets (Adak et al., 2022; Alsagheer et al., 2022), discussing automated approaches to counterspeech classification (Alsagheer et al., 2022) and generation (Chaudhary et al., 2021; Alsagheer et al., 2022), and work focusing on system evaluation (Alsagheer et al., 2022). However, NLP work from computer sciences is not typically informed by important insights from the social sciences, including the key roles of intergroup dynamics, the social context in which counterspeech is employed, and the mode of persuasion by which counterspeech operates. Taking an interdisciplinary approach, we join work from the computer and social sciences.

3 Review Methodology

Taking a multi-disciplinary perspective, we systematically review work on counterspeech from computer science and the social sciences published in the past ten years. To ensure broad coverage and to conduct a reproducible review, we follow the systematic methodology of Moher et al. (2009). The search and inclusion process is shown in Figure 2.

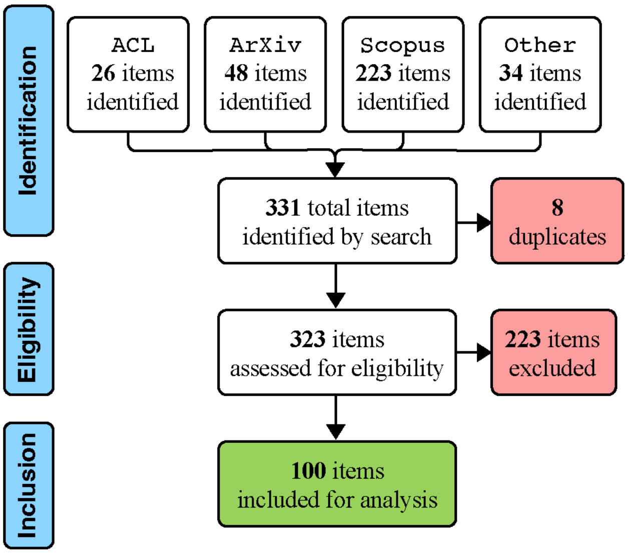


Figure 2: Flow diagram showing the identification, eligibility screening, and inclusion phases of the selection of items analysed in this review.

We used keyword terms related to counterspeech to search three key databases (ACL Anthology, ArXiv, and Scopus) that together offer a broad coverage of our target literature. We included the search terms ‘counter-speech’, ‘counter-narratives’, ‘counter-terrorism’, ‘counter-aggression’, ‘counter-hate’, ‘counter speech’, ‘counter narrative’, ‘countering online hate speech’, ‘counter hate speech’, and ‘counter-hate speech’. We also included 34 publications that we had identified previously from other sources, but that were not returned by keyword search due to not including relevant keywords or not being indexed in the target search repositories. The search covers the data within the period between 2005 and 2023. Of the returned results, we include all publications that concern (1) analysis of the use and effectiveness of interventions against hateful or abusive language online, (2) characteristics of counterspeech users or recipients, or (3) data and/or implementation designed for counterspeech (e.g., counterspeech classification or generation). These inclusion criteria were applied by two of the authors. Following this process, we include 100 papers for analysis in this review. Each of the papers was read by at least one of the co-authors of the article.

Our review is divided into several sections (the results of which are presented sequentially below). First, we examine definitional characteristics of counterspeech, looking at how the term itself is defined, how different taxonomies have been created to classify different types of counterspeech, and the different potential purposes attributed to it. Based on the definitional characteristics, we examine studies that have looked at the impact of counterspeech, discussing the different analytical designs employed and analysing evidence of the results. Following this, we discuss computational approaches to counterspeech, focusing in particular on both detection and generation. Finally, we examine ethical issues in the domain of counterspeech, and also speculate about future perspectives and directions in the field.

4 Defining counterspeech

Counterspeech is multifaceted and can be characterised in several different ways. In Table 1 we outline a framework for describing and designing counterspeech, covering who (speaker) sends what kinds of messages (strategies) to whom (recipients), and for what purpose (purpose). Using this structure, we summarise how counterspeech has typically been categorised in past studies.

Most studies in the field use one of three main terms: *counterspeech*, *counter-narratives* (Reynolds and Tuck, 2016; Carthy and Sarma, 2021; Tuck and Silverman, 2016; Iqbal et al., 2019) and *hope speech* (Snyder et al., 2018). These three terms broadly refer to a similar concept: content that challenges and rebuts hateful discourse and propaganda (Saltman and Russell, 2014; Bartlett and Krasodomski-Jones, 2015; Benesch et al., 2016; Saltman et al., 2021; Garland et al., 2022) using non-aggressive speech (Benesch et al., 2016; Reynolds and Tuck, 2016; Schieb and Preuss, 2016). There are some differences between the terms. Ferguson (2016) considers counter-narratives as intentional strategic communication within a political, policy, or military context. Additionally, the term counter-narrative also refers to narratives that challenge a much broader view or category such as forms of education, propaganda, and public information (Benesch et al., 2016). Such counter-narratives are often discussed in the context of the prevention of violent extremism. Hope speech, meanwhile, could be seen as a particular type of counterspeech: it promotes positive engagement in online discourse to lessen the consequences of abuse, and places a particular emphasis on delivering optimism, resilience, and the values of equality, diversity and inclusion (Chakravarthi, 2022). In this paper, we review work that relates to all of these three concepts, and largely make use of the catch-all term counterspeech, while acknowledging the slight differences between the concepts.

4.1 Classifying counterspeech

Researchers have identified a variety of different types of counterspeech. Here, we outline four main ways in which counterspeech can vary, in terms of the identity of the counterspeaker, the strategies employed, the recipient of the counterspeech and the purpose of counterspeech.

Counterspeakers (who) Psychological studies show that the identity of a speaker plays a key role in how large an audience their message reaches and how persuasive the message is. Common crucial factors include group identity (such as race, religion, and nationality), level of influence, and socioeconomic status. For instance, counterspeech provided by users with large numbers of followers and from an in-group member is more likely to lead to changes in the behaviour of perpetrators of hate (Munger, 2017).

Some studies characterise individuals who use counterspeech and suggest that these users exhibit different characteristics and interests than users who spread hate (Mathew et al., 2018, 2019; Buerger, 2021b). Through lexical, linguistic and psycholinguistic analysis of users who generate hate speech or counterspeech on Twitter, Mathew et al. (2018) find that counterspeakers are higher in agreeableness, displaying traits such as altruism, modesty, and sympathy, and display higher levels of self-discipline and conscientiousness. Possibly driven by a motive to help combat hate speech, counterspeakers tend to use words related to government, law, leadership, pride, and religion. Regarding the impact of being a counterspeaker, in an ethnographic study, members of a counterspeech campaign reported feeling more courageous and keen to engage in challenging discussions after expressing opinions publicly (Buerger, 2021b).

Strategies (how) Counterspeech can take many forms. Benesch et al. (2016) first identify eight types of counterspeech used on Twitter: (1) *presentation of facts*, (2) *pointing out hypocrisy or contradiction*, (3) *warning of consequences*, (4) *affiliation* [i.e. establishing an emotional bond with the perpetrators or targets of hate], (5) *denouncing*, (6) *humour/sarcasm*, (7) *tone* [a tendency or style adopted for communication, e.g., empathetic and hostile], and (8) *use of media*. Based on this taxonomy, follow-up studies on counterspeech make minor modifications to cover strategies in a broader scope. Mathew et al. (2018) analyzed and classified counterspeech on Twitter, taking Benesch et al. (2016)'s taxonomy but dropping *the use of media* and adding *hostile language* and *positive tone*, which replaces general strategy *tone*. Similarly, Mathew et al. (2019) collected and annotated counterspeech comments from Youtube, adopting Benesch et al. (2016)'s taxonomy but excluding *tone* and adding *positive tone*, *hostile language* and *miscellaneous*. Chung et al. (2019) collaborated with NGOs to collect manually written counterspeech. For data annotation, they followed the taxonomies pro-

Aspects	Description
Speaker	Who is the counterspeaker? What is the social identity and status of the counterspeaker?
Strategy	Which linguistic and rhetorical methods are used in the counterspeech? Which emotions or attitudes are expressed towards the hateful content?
Recipient	Who is the target audience? Are they hate speakers, targets of hate, or bystanders?
Purpose	What is the aim of disseminating counterspeech?

Table 1: Framework for describing and designing counterspeech.

vided by Benesch et al. (2016) and Mathew et al. (2019), while adding *counter question* and discarding *the use of media*. Counterspeech examples for each strategy are provided in Table 2.

Counterspeech recipients (whom) Depending on the purpose of the counterspeech, the target audience may be perpetrators, victims or bystanders (see Figure 1). Identifying the appropriate target audience or 'Movable Middle' is crucial to maximise the efficacy of counterspeech. Movable middle refers to individuals who do not yet hold firm opinions on a topic and can hence be potentially open to persuasion. They are also receptive to arguments and more willing to listen. These individuals often serve as ideal recipients of messages addressing social issues such as vaccination hesitancy (Litaker et al., 2022). In the context of counterspeech, previous studies show that a small group of counterspeakers can shape online discussion when the audience holds moderate views (Schieb and Preuss, 2016; Buerger, 2021b).

Wright et al. (2017) group counterspeech acts into four categories based on the number of people involved in the discussion: *one-to-one*, *one-to-many*, *many-to-one*, or *many-to-many*. Some successful cases where counterspeech induces favourable changes in the discourse happen in a one-to-one discussion. This allows for dedicated opinion exchange over an ideology, which in some cases even yields long-lasting changes in beliefs. The use of hashtags is a good example of one-to-many and many-to-many interaction where conversations surge quickly (Benesch et al., 2016; Wright et al., 2017). For instance, Twitter users often include hashtags to express support (e.g., #BlackLivesMatter) or disagreement with haters (e.g., #StopHate) to demonstrate their perspective.

The purpose of counterspeech Hateful language online can serve to reinforce prejudice (Citron and Norton, 2011), encourage further division, promote power of the ingroup, sway political votes, provoke or justify offline violence, and psychologically damage targets of hate (Jay, 2009). Just as the effects of hate are wide-ranging, counterspeech may be used to fulfil a variety of purposes.

- ***Changing the attitudes and behaviours of perpetrators*** In directly challenging hateful language, one key aim of counterspeech can be to change the attitudes of the perpetrators of hate themselves. The strategy here is often to persuade the perpetrator that their attitudes are mistaken or unacceptable, and to deconstruct, discredit or delegitimise extremist narratives and propaganda (Reynolds and Tuck, 2016). Counterspeech aimed at changing the attitudes of spreaders of hate may address the hate speaker directly, countering claims with facts or by employing empathy and affiliation. Challenging attitudes is often seen as a stepping stone to altering behaviours (Stroebe, 2008). In attempting to change the minds of perpetrators, counterspeakers ultimately hope to discourage associated behaviours such as sharing such content again in the future or showing support for other hateful content (i.e., stopping the spread of hate). In changing the minds of perpetrators, counterspeakers may also hope to prevent them from engaging in more extreme behaviours such as offline violence.

- ***Changing the attitudes and behaviours of bystanders*** More commonly, counterspeech is initiated with the intention of reaching the wider audience of bystanders rather than perpetrators of hate themselves (Buerger, 2022). These bystanders are not (at least yet) generating hateful language themselves, but rather are people exposed to hateful content either incidentally or by active engagement. Here, counterspeakers hope to persuade bystanders that the hateful content is wrong or unacceptable, again by deconstructing and delegitimising the hateful narrative. The strategy here may be to offer facts, point out hypocrisy, denounce the content, or use humour to discredit the speaker. Additionally, counterspeakers will often invoke empathy for targets of hate. In preventing bystanders from forming attitudes and opinions in line with the hateful narrative, counterspeakers hope to mitigate further intergroup division and related behaviours such as support for or engagement with additional abuse or physical violence. Counterspeakers may also hope to encourage others to generate rebutals and rally support for victims (Benesch, 2014a), bringing positive changes in online discourse.

- ***Showing support for targets of hate*** A third key way in which counterspeech functions is to show

Strategy	Example
Facts	Actually, studies show that on the whole migrants contribute more to public finances than they take out, see this article for example.
Hypocrisy	Immigrants stealing British resources? A bit rich given how much was stolen from colonies by the British Empire.
Consequences	Spreading hateful content is illegal. Police will knock on your door.
Affiliation	As a British national, I know life is hard here right now. But I assure you that your unemployment is not the fault of immigrants.
Denouncing	Stop with the racist and derogatory slurs. It's unacceptable to talk this way.
Counter questions	Do you have a problem with all immigrants or only ones from lower income countries? Are you suggesting we have enough qualified and willing British born workers to fill all the jobs?
Humour	You should think about how the Spanish feel next time you go on holiday to Costa Del Sol (laughing emoji)?
Positive tone	Immigrants strengthen UK society in so many ways - greater diversity, skillsets and innovation to name a few! And no way our NHS could function without the immigrant workforce.

Table 2: Synthetic examples of different counterspeech strategies in response to an example of abuse against immigrants. Here the abuse example is: 'Immigrants are invading and stealing our resources'.

support directly to targets of hate. Online abuse can psychologically damage the wellbeing of targets and leave them feeling fearful, threatened, and even in doubt of their physical safety (Benesch, 2014b; Leader Maynard and Benesch, 2016; Saha et al., 2019; Siegel, 2020). By challenging such abuse, counterspeakers can offer support to targets and encourage bystanders to do the same (Buerger, 2021b). This support aims to alleviate negative emotion brought on by hate by demonstrating to targets that they are not alone and that many people do not hold the attitudes of the perpetrator. Here the particular strategies may be to denounce the hate and express positive sentiment towards the target group. Intergroup solidarity may in turn reduce retaliated antagonism.

5 The Impact of Counterspeech

While we have delineated the characteristics of counterspeech, its concrete effects on harm mitigation remain debated. The methods applied for evaluating the effectiveness of counterspeech vary considerably across studies in the field. In this section we provide an evidence-based analysis of counterspeech's efficacy, examining how it is used in real-life scenarios and its influence based on eight aspects.

Research design A wide range of methodologies have been adopted to assess the impact of counterspeech on hate mitigation, including observational studies (Ernst et al., 2017; Stroud and Cox, 2018; Garland et al., 2022), experimental (Munger, 2017; Obermaier et al., 2021; Hangartner et al., 2021) and quasi-experimental designs (Bilewicz et al., 2021). In observational studies, investigators typically assess the relationship between exposure to counterspeech and outcome variables of interest without any experimental manipulation. For instance, a longitudinal study of German political conversations on Twitter examined the interplay between organized hate and counterspeech groups (Garland et al., 2022). There is also an ethnographic study interviewing counterspeakers on Facebook to understand external and internal practices for collectively intervening in hateful comments, such as how to build effective counterspeech action and keep counterspeakers engaged (Buerger, 2021b). For experimental and quasi-experimental designs, both aim at estimating the causal effects of exposure to different kinds of counterspeech on outcome variables in comparison with controls (no exposure to counterspeech).

Languages and countries In the reviewed work, the impact of counterspeech is investigated in five different languages across nine countries. Notably, experiments are focused on counterspeech used in Indo-European languages such as English (USA, UK, Canada and Ireland), German (Germany), Urdu (Pakistan) and Swedish (Sweden). Only two studies are dedicated to Afro-Asiatic languages, Arabic (Egypt and Iraq). We did not find research dedicated to other language families, suggesting that the language coverage of counterspeech studies is still low.

Platforms Most experiments were conducted on text-based social media platforms, such as eight on Twitter (Benesch et al., 2016; Reynolds and Tuck, 2016; Silverman et al., 2016; Stroud and Cox, 2018; Munger, 2017; Hangartner et al., 2021; Poole et al., 2021; Garland et al., 2022), six on Facebook (Reynolds and Tuck, 2016; Silverman et al., 2016; Schieb and Preuss, 2016;

Leonhard et al., 2018; Saltman et al., 2021; Buerger, 2021b), and one on Reddit (Bilewicz et al., 2021), as well as image-based online spaces, such as three on Youtube (Reynolds and Tuck, 2016; Silverman et al., 2016; Ernst et al., 2017) and one on Instagram (Stroud and Cox, 2018). Often, the counterspeech interventions are directly monitored on such platforms, but in some cases, fictitious platforms are created in order to mimic online social activity under a controlled environment (Obermaier et al., 2021; Carthy and Sarma, 2021; Bélanger et al., 2020). There are three studies analysing the impact of counterspeech across multiple platforms (Reynolds and Tuck, 2016; Silverman et al., 2016; Stroud and Cox, 2018).

Twitter and Facebook are widely used for measuring the effects of counterspeech, with eight and six experiments respectively. For Twitter, this can be explained by its easily accessible API (even if at the time of writing continued research access to the API was in doubt). Similarly, because of difficulties in gathering data, Schieb and Preuss (2016) resort to developing an agent-based computational model for simulating hate mitigation with counterspeech on Facebook. It is worth highlighting that none of the studies we reviewed had investigated recently popular mainstream platforms, such as Tiktok, Weibo, Telegram, and Discord.

The target of hate speech Abusive speech can be addressed towards many different potential targets, and each individual hate phenomenon may require different response strategies for maximum effectiveness. Existing studies have evaluated the effectiveness of counterspeech on several hate phenomena, with Islamophobia, Islamic extremism, and racism being the most commonly addressed, while hate against LGBTQ+ community and immigrants being the least studied. In these studies, abusive content is typically identified based on two strategies - hateful keyword matches (Hangartner et al., 2021; Bilewicz et al., 2021), or user accounts (e.g., content produced by known hate speakers) (Garland et al., 2022).

Types of interventions A wide range of methods are exploited to design and surface counterspeech messages to a target audience. We broadly categorise these methods based on modality and approach to creation. Counter speech is generally conveyed in text (Bélanger et al., 2020; Hangartner et al., 2021; Poole et al., 2021) or video mode (Ernst et al., 2017; Saltman et al., 2021; Carthy and Sarma, 2021). In both cases, counterspeech materials can be created in three different ways: written by experimenters as stimuli (Obermaier et al., 2021; Carthy and Sarma, 2021), as well as written by individuals or campaigns that are collected from social media platforms (Benesch et al., 2016; Garland et al., 2022; Buerger, 2021b). We also found one study integrating counterspeech messages in media such as films, TV dramas and movies (Iqbal et al., 2019).

Counterspeech strategies Following the strategies summarised in Section 4.1, commonly used counterspeech strategies include facts (Buerger, 2021b; Obermaier et al., 2021), denouncing (Stroud and Cox, 2018; Saltman et al., 2021), counter-questions (Silverman et al., 2016; Reynolds and Tuck, 2016; Saltman et al., 2021), and a specific tone (humour or empathy) (Reynolds and Tuck, 2016; Munger, 2017; Hangartner et al., 2021; Saltman et al., 2021). There are more fine-grained tactics for designing counterspeech in social science experiments. According to psychological studies, the use of social norms can reduce aggression and is closely related to legal regulation in society (Bilewicz et al., 2021). This tactic was tested in an intervention study where participants were exposed to counterspeech with one of the inducements of empathy, descriptive norms (e.g., *Let's try to express our points without hurtful language*) and prescriptive norms (e.g., *Hey, this discussion could be more enjoyable for all if we would treat each other with respect.*) (Bilewicz et al., 2021). Bélanger et al. (2020) designed counterspeech based on substances rather than tactics, varying three different narratives: (1) social (seeking to establish a better society), (2) political (bringing a new world order through a global caliphate), and (3) religious (legitimising violence based on religious purposes). Considering broader counterspeech components, a few organisations further focus on challenging ideology (e.g., far-right and Islamist extremist recruitment narratives), rather than deradicalising individuals (Silverman et al., 2016; Saltman et al., 2021). Counterspeech drawing from personal stories in a reflective or sentimental tone is also considered as it can resonate better with target audiences (Silverman et al., 2016). In addition to neutral or positive counterspeech, radical approaches are taken by counter-objecting, degrading or shaming perpetrators in public for unsolicited harmful content (Stroud and Cox, 2018; Obermaier et al., 2021).

Types of evaluation metrics Based on Reynolds and Tuck (2016)'s counterspeech *Handbook*, we identified the following three types of metrics used by the authors of the papers to evaluate the effectiveness of counterspeech interventions: social impact, behavioural change, and attitude change measures.

• ***Social impact metrics*** are (usually automated) measurements of how subjects interact with counterspeech online. Such measures include, bounce rate, exit rate,[4]

[4] Bounce rate is the number of users who leave a website without clicking past the landing page; exit rate measures how many people leave the site from a given section (Reynolds and Tuck, 2016).

geo-location analysis and the numbers of likes, views, and shares that posts receive (Garland et al., 2020; Hangartner et al., 2021; Poole et al., 2021; Reynolds and Tuck, 2016; Leonhard et al., 2018; Saltman et al., 2021; Silverman et al., 2016). For example, for one of their experiments, Saltman et al. (2021) measure the 'click-through rates' of Facebook users redirected from hateful to counterspeech materials, while Hangartner et al. (2021) measure retweets and deletions (in addition to behavioural change measures).

Social impact measures are also applied to synthetic data by Schieb and Preuss (2016), who measure the 'likes' of their (simulated) participants as hate and counterspeech propagate through a network (as well as applying behavioural metrics). Taking a more distant, long-term view, Iqbal et al. (2019) cite Egypt's overall success at countering radicalisation with counterspeech campaigns by comparing its position on the Global Terrorism Index with that of Pakistan.

While the majority of these measurements are automated, Leonhard et al. (2018) use survey questions to examine participants willingness to intervene against hate speech depending on the severity of the hate, the number of bystanders, and the reactions of others. Unlike the survey-based approaches described below, they do not consider *changes* in attitude. In addition, Buerger (2021b) assess the success of the #jagärhär counterspeech campaign (#iamhere in English, a Sweden-based collective effort that has been applied in more than 16 countries) based on the extent to which it has facilitated the emergence of alternative perspectives.

- ***Behavioural change measures*** reveal whether subjects change their observable behaviour towards victims before and after exposure to counterspeech, for example in the tone of their language as measured with sentiment analysis.

For instance, Hangartner et al. (2021) conduct sentiment analysis to determine the behaviour of previously xenophobic accounts after treatment with counterspeech, Bilewicz et al. (2021) measure levels of verbal aggression before and after interventions, and Garland et al. (2020) assess the proportion of hate speech in online discourse before and after the intervention of an organised counterspeech group. Other such measures are those of Saltman et al. (2021), who compare the number of times users that violate Facebook policies before and after exposure to counterspeech, and Munger (2017), who examine the likelihood of Twitter users continuing to use racial slurs following sanctions by counterspeakers of varying status and demographics. And in a network simulation experiment, Schieb and Preuss (2016) measure the effect of positive or negative (synthetic) posts on (synthetic) user behaviour.

- ***Attitude change measures*** are used to assess whether people (hate/counter speakers or bystanders) change their underlying attitudes or intentions through non-automated methods such as interviews, surveys, focus groups, or qualitative content analysis.

For potential hate speech perpetrators, Carthy and Sarma (2021) use psychological testing to measure the extent to which participants legitimized violence after exposure to differing counterspeech strategies, Bélanger et al. (2020) compare support for ISIS and other factors using in participants exposed to differing counterspeech strategies and a control group, and Ernst et al. (2017) code user comments on hate and counterspeech videos to perform qualitative content analysis of users' attitudes.

For bystanders that may be potential counterspeakers, Obermaier et al. (2021) use a survey to examine whether counterspeech leads to increased intentions to intervene. And for those already engaged in counterspeech, Buerger (2021b) conduct interviews with members of an organised group to reveal their perceptions of the efficacy of their interventions.

Effectiveness Owing to the variation in experimental setups, aims, and evaluation methods of the counterspeech efforts we review, it is not straightforward to compare their levels of success. Indeed, several of the studies concern broad long-term goals that cannot be easily evaluated at all (e.g. Reynolds and Tuck, 2016; Silverman et al., 2016) or provide only anecdotal evidence (e.g. Benesch et al., 2016; Stroud and Cox, 2018; Buerger, 2021b).

Beyond this, evidence of successful counterspeech forms a complex picture. For example, Garland et al. (2022) show that organised counterspeech is effective, but can produce backfire effects and actually attract more hate speech in some circumstances. They also show that these dynamics can alter surrounding societal events—although they do not make causal claims for this. Similarly, Ernst et al. (2017) find mixed results, with counterspeech encouraging discussion about hate phenomena and targets in some cases, but also leading to increases in hateful comments. However, Silverman et al. (2016) suggest that even such confrontational exchanges can be viewed as positive signs of engagement.

There is some evidence for the comparative efficacy of different counterspeech strategies. Bilewicz et al. (2021) find that three of their intervention types ('disapproval', 'abstract norm', 'empathy') are effective in reducing verbal violence when compared with no intervention at all. Here, empathy had the weakest effect, which they put down to the empathetic messages being specific to particular behaviours, limiting their capacity to modify aggression towards wider targets. Hangartner et al. (2021) also found that empathy-based counterspeech can consistently reduce hate speech, although this effect is small. And Carthy and Sarma (2021) found that counterspeech that seeks to correct false information in the hate speech actually leads to higher levels of violence legitimisation, while having participants actively counter terrorist rhetoric themselves ('Tailored Counter-Narrative') was the most effective strategy to reduce this. They found counterspeech to be more effective on participants that are already predisposed to cognitive reflection. However, focusing on the effect of factual correction on the victims rather than perpetrators of hate speech, Obermaier et al. (2021) found it to be effective in providing support and preventing them from hating back and therefore widening the gap between groups.

There is also some evidence that the numbers of the different actors involved in a counterspeech exchange can affect an intervention's success. Schieb and Preuss (2016) find that counterspeech can impact the online behaviour of (simulated) bystanders, with the effectiveness strongly influenced by the proportions of hate and counter speakers and neutral bystanders. According to their model, a small number of counterspeakers can be effective against smaller numbers of hate speakers in the presence of larger numbers of people lacking strong opinions. Saltman et al. (2021) found their counterspeech strategies to be effective only for higher risk individuals within the target populations, although they did not see any of the potential negative effects of counterspeech (such as increased radicalisation) reported elsewhere.

Focusing on who in particular delivers counterspeech, Munger (2017) finds that success of counterspeech depends on the identity and status of the speaker. However, with only a small positive effect, Bélanger et al. (2020) found that the content of counterspeech was more important than the source. And Garland et al. (2022) found that, while organised counterspeech can be effective, the efforts of individuals can lead to increases in hate speech. In Buerger (2021b), members of #jagärhär claim that their counterspeech interventions were successful in making space for alternative viewpoints to hate speech.

6 Computational Approaches to Counterspeech

In this section, we switch the focus to look at NLP literature on counterspeech emerging from the field of computer science. We tackle three subjects in particular: the datasets being used in these studies, approaches to counterspeech detection, and approaches to counterspeech generation.

6.1 Counterspeech Datasets

Collection strategies Approaches for counterspeech collection focus on gathering two different kinds of datasets: spontaneously produced comments crawled from social media platforms, and deliberately created responses aiming to contrast hate speech. In the first case, content is retrieved based on keywords/hashtags related to targets of interest (Mathew et al., 2018; Vidgen et al., 2020; He et al., 2022; Vidgen et al., 2021) or from pre-defined counterspeech accounts (Garland et al., 2020). In principle, due to the easily accessible API required for data retrieval, the majority of datasets are collected from social media platforms including Twitter (Mathew et al., 2018; Procter et al., 2019; Garland et al., 2020; Kennedy et al., 2020; Vidgen et al., 2020; He et al., 2022; Goffredo et al., 2022; Toliyat et al., 2022; Lin et al., 2022), and only a few are retrieved from Youtube (Mathew et al., 2019; Kennedy et al., 2020; Priyadharshini et al., 2022) and Reddit (Kennedy et al., 2020; Vidgen et al., 2021; Lee et al., 2022; Yu et al., 2022), respectively (though again it is worth noting that at the time of writing the Twitter API was starting to become a lot less accessible). To find the best strategy for collecting online content, Möhle et al. (2023) compare the keywords-matching method with automated filtering using a multilingual model fine-tuned on English data for German counterspeech collection. They found neither strategy helped curate significantly more counterspeech compared to a random sampling baseline.

In the second category, counterspeech is written by crowd workers (Qian et al., 2019) or operators expert in counterspeech writing (Chung et al., 2019, 2021c). While such an approach is expected to offer relatively controlled and tailored responses, writing counterspeech from scratch is time-consuming and requires human effort. To address this issue, advanced generative language models are adopted to automatically produce counterspeech (Tekiroğlu et al., 2020; Fanton et al., 2021; Bonaldi et al., 2022), as we will discuss further below.

Granularity and languages Regarding granularity of taxonomies, most existing datasets provide binary annotation (counterspeech/non-counterspeech) (Garland et al., 2020; Vidgen et al., 2020; He et al., 2022; Vidgen et al., 2021), while three datasets feature annotations of the types of counterspeech (Mathew et al., 2018, 2019; Chung et al., 2019). Recently, Yu et al. (2023) propose a taxonomy that distinguishes the target of counterspeech (i.e. whether the counterspeech addresses the hateful content or the author of the hateful comment) and identifies the argument components in the counterspeech (i.e. logical arguments and appealing to emotion). In terms of hate incidents, datasets are available for several hate phenomena such as Islamophobia (Chung et al., 2019) and East Asian prejudice during the COVID-19 pandemic (Vidgen et al., 2020; He et al., 2022). The aforementioned datasets are mostly collected and analyzed at the level of individual text, not at discourse or conversations (e.g., multi-turn dialogues (Bonaldi et al., 2022)). Most of the datasets are in English, while only a few target multilinguality, including Italian (Chung et al., 2019; Goffredo et al., 2022), French (Chung et al., 2019), Spanish (Vallecillo-Rodríguez et al., 2023), German (Garland et al., 2020; Möhle et al., 2023), and Tamil (Priyadharshini et al., 2022).

6.2 Approaches to Counterspeech Detection

Previous work on counterspeech detection has focused on binary classification (i.e. whether a text is counterspeech or not) (Vidgen et al., 2020; Garland et al., 2022; He et al., 2022) or identifying the types of counterspeech as a multi-label task (Mathew et al., 2018; Garland et al., 2020; Chung et al., 2021a; Goffredo et al., 2022). Automated classifiers are developed to analyse large-scale social interactions of abuse and counterspeech addressing topics such as political discourse (Garland et al., 2022) and multi-hate targets (Mathew et al., 2018). Moving beyond monolingual study, Chung et al. (2021a) evaluate the performance of pre-trained language models for categorising counterspeech strategy for English, Italian and French in monolingual, multilingual and cross-lingual scenarios.

6.3 Approaches to Counterspeech Generation

Various methodologies have been put forward for the automation of counterspeech generation (Qian et al., 2019), addressing various aspects including the efficacy of a hate countering platform (Chung et al., 2021b), informativeness (Chung et al., 2021c), multilinguality (Chung et al., 2020), politeness (Saha et al., 2022), and grammaticality and diversity (Zhu and Bhat, 2021). These methods are generally centred on transformer-based large language models (e.g., GPT-2 (Radford et al., 2019)). By testing various decoding mechanisms using multiple language models, Tekiroğlu et al. (2022) find that autoregressive models combined with stochastic decoding yield the optimal counterspeech generation. In addition to tackling hate speech, there are studies investigating automatic counterspeech generation to respond to trolls (Lee et al., 2022) and microagressions (Ashida and Komachi, 2022).

Evaluation of counterspeech generation Assessing counter speech generation is complex and challenging due to the lack of clear evaluation criteria and robust evaluation techniques.

Previous work evaluates the performance of counterspeech systems via two aspects: automatic metrics and human evaluation. Automatic metrics, generally, evaluate the generation quality based on criteria such as linguistic surface (Papineni et al., 2002; Lin, 2004), novelty (Wang and Wan, 2018), and repetitiveness (Bertoldi et al., 2013; Cettolo et al., 2014). Despite being scalable, these metrics are uninterpretable and can only infer model performance according to references provided (e.g., dependent heavily on exact word usage and word order) and gathering an exhaustive list of all appropriate counterspeech is not feasible. For this reason, such metrics cannot properly capture model performance, particularly for open-ended tasks (Liu et al., 2016; Novikova et al., 2017) including counterspeech generation. As a result, human evaluation is heavily employed based on aspects such as suitableness, grammatical accuracy and relevance (Chung et al., 2021c; Zhu and Bhat, 2021). Despite being trusted and high-performing, human evaluation has inherent limitations such as being costly, difficult (e.g., evaluator biases and question formatting), and time-consuming (both in terms of evaluation and moderator training), and can be inconsistent and inflict psychological harm on the moderators.

The effectiveness of counterspeech generations should be also carefully investigated 'in-the-wild' to understand its social media impact, reach of content, and the dynamics of hateful content and counterspeech (see Section 5). This line of research is limited. The closest work to this research space is by Zheng et al. (2023) that identifies the characteristics of good counterspeech in terms of the quality and effectiveness and user preference for machine-generated counterspeech through a survey. Based on 29 subjects (i.e. bystanders) evaluating 60 pseudo-threads on Twitter (at the time of experiments), they conclude that clear and direct responses with thorough explanations are mostly preferred by users.

Potentials and limits of existing generative models We believe that in some circumstances counterspeech may be a more appropriate tool than content moderation in fighting hate speech as it can depolarise discourse and show support to victims. However, automatic counterspeech generation is a relatively new research area. Recent progress in natural language processing has made large language models a popular vehicle for generating fluent counterspeech. However, counterspeech generation currently faces several challenges that may constrain the development of efficient models and hinder the deployment of hate intervention tools. Similar to the use of machine translation and email writing tools, we advocate that counterspeech generation tools should be deployed as suggestion tools to assist in hate countering activity (Chung et al., 2021c,b).

- ***Faithfulness/Factuality in generation*** Language models are repeatedly reported to produce plausible and convincing but not necessarily faithful/factual statements (Solaiman et al., 2019; Zellers et al., 2019; Chung et al., 2021c). We refer to faithfulness as being consistent and truthful in adherence to the given source (i.e. model inputs) (Ji et al., 2023). Such unfaithful/non-factual generation is particularly intolerable for counterspeech generation as it can create unwanted consequences or elicit hatred. Many attempts have been made to mitigate this issue (Ji et al., 2023) such as correcting unfaithful data (Nie et al., 2019) and measuring faithfulness of generated outputs (Dušek and Kasner, 2020; Zhou et al., 2021). For the task of counterspeech generation, Chung et al. (2021c) present the first knowledge-bound generation pipeline consisting of a knowledge retrieval module that retrieves relevant knowledge to the context of hate speech and a generation module that generates a counterspeech response. Following this approach, Jiang et al. (2023) employ a retrieval-augmented unsupervised generation method that refines retrieved knowledge based on stance consistency and semantic overlap for hate speech and allows for generation without gold-standard data. In a similar vein, Furman et al. (2023) prompt large language models with argumentative information in hate speech to enhance the quality of counterspeech generation and show that this approach is especially beneficial for low-resource scenarios. To facilitate reliable counterspeech generation applications, we encourage reporting the faithfulness/factuality of models.

- ***Toxic degeneration and debiasing*** Language models can also induce unintendedly biased and/or toxic content, regardless of whether explicit prompts are used (Dinan et al., 2022). In the use case of counterspeech generation, this can result in harm to victims and bystanders as well as risking provoking perpetrators into further abusive behaviour. This issue has been mitigated by two approaches: data and modelling. The data approach aims at creating proper datasets for fairness by removing undesired and biased content (Blodgett et al., 2020; Raffel et al., 2020). The modelling approach focuses on controllable generation techniques that, for instance, employ humans for post-editing (Tekiroğlu et al., 2020) and detoxification techniques (Gehman et al., 2020). Another line of research emphasises that implicit stereotypical beliefs or biases from hateful content should be addressed in counterspeech generation (Mun et al., 2023; Akazawa et al., 2023). For instance, Akazawa et al. (2023) tune large language models to infer implicit biases from hate speech and found that such extra information helps improve generation quality.

- ***Diversity, Generalisation and Specialisation*** With the rise of online hate, models that can generalize across domains would help produce counterspeech involving new topics and events, while it may come with the cost of losing specificity. Generalisable methods can ameliorate the time and manual effort required for collecting and annotating data. However, as discussed in Section 5, counterspeech is multifaceted and contextualised. For instance, abuse against women can often be expressed in a more subtle form as microaggressions. Specific and diverse responses to hateful or prejudiced language are often preferred as they can provide coherent discourse relations and potential connection with personal events (Finnegan et al., 2015). In a user study comparing model-generated and human-written counterspeech, Mun et al. (2023) show that humans prefer and use more specific strategies targeting stereotypical statements when countering hate while models tend to produce less convincing arguments according to annotators. To produce more specific responses, Hassan and Alikhani (2023) show that grounding generation in context using discourse-augmented prompting strategies results in contextual, diverse and accurate counterspeech. Similarly, Gupta et al. (2023) propose to guide generation based on five intents (informative, question, denouncing, humour, and positive) for generating diverse counterspeech. To address the generalisation capabilities of large language models for counterspeech generation, Bonaldi et al. (2023) introduce attention-based regularisation techniques that help contextualise token representations (i.e. include broader hate speech context) and guide models to focus on specific attention distributions (e.g. use words related to minority targets). There may not be a one-size-fits-all solution. Overall, model generalisability is still challenging (Fortuna et al., 2021; Yin and Zubiaga, 2021), and can have potential limitations (Conneau et al., 2020; Berend, 2022). Finding the right trade-off between generalisation and specialisation is key.

7 Future Perspectives

Of the many promising abuse intervention experiments that we review, results are not always consistent, demonstrating weak claims or limited success (applicable only to certain settings). Possible reasons include short-term experiments, small sample sizes and non-standardised experimental designs. To improve this, effective interventions should come with the characteristics of scalability, durability, reliability, and specificity. In this section, we highlight key distinctions and overlaps across areas that have and have not been explored in social sciences and computer science, discuss ethical issues related to evaluating counterspeech in real-life settings and automating the task of counterspeech generation, and identify best practices for future research.

Distinctions and overlaps across areas By recognizing the commonalities and differences between social sciences and computer science, we pinpoint the unique contributions of each discipline and encourage interdisciplinary collaborations to address complex societal challenges and better understand human behaviour with the help of computational systems.

- ***Terminological clarity*** Throughout the counterspeech literature, terminology is used inconsistently. Terms such as counterspeech and counter-narratives are often used interchangeably or used to refer to similar concepts. In social science, counterspeech is used to refer to content that disagrees with abusive discourses and counter-narratives often entail criticism of an ideology with logical reasoning. As a result, counter-narrative stimuli designed in social experiments are generally long form (Bélanger et al., 2020). In computer science on the other hand, the distinctions between counterspeech and counter-narratives have been vague, and training data is generally short form (while this may be bound by character limit on social media platforms). For instance, short and generic responses such as '*How can you say that about a faith of 1.6 billion people?*' can be commonly found in counter-narrative datasets (Chung et al., 2019).

- ***The focus of evaluation*** Social scientists and counterspeech practitioners generally attempt to understand and assess the impact of counterspeech on reducing harms (e.g., which strategies are effective and public perception towards counterspeech), whereas computer scientists focus more on technical exploration of automated systems and testing their performance in producing counterspeech (e.g., comparing system outputs with a pre-established ground truth or supposedly ideal output). One commonality between the social science and computer science studies is that most findings are drawn from controlled and small-scale studies. Applying interventions to real-world scenarios is a critical next step.

- ***Datasets*** Dataset creation is an important component in computer science for developing machine learning models for generating counterspeech, while such contributions are less commonly considered in social sciences which rely on experiments using hand-crafted stimuli and one-time analyses of their effectiveness.

- ***Scope of research*** We observe that, while computer scientists have focused on responses to abusive language and hate speech, social science studies address a wider range of phenomena, in particular radicalisation and terrorist extremism. It can be difficult to measure the effectiveness of counterspeech in challenging these over the short term, leading to some of the differences in evaluation metrics across disciplines.

- ***Lack of standardised methodologies*** A variety of methodologies have been adopted in the literature, making comparisons across studies difficult. Without standardised evaluations, it is difficult to situate the results and draw robust findings.

Ethical Issues, Risks and Challenges of Conducting Counterspeech Studies Effective evaluation of counterspeech not only identifies users who may need help, but also safeguards human rights and reinforces a stronger sense of responsibility in the community. This discussion is based on the authors' opinion and not stemming from the review.

- ***Evaluating counterspeech in real-life settings*** Conducting the evaluation of counterspeech in real-world scenarios appears to provide a proactive and quick overview of its performance on hate mitigation. Nevertheless, the best ways to approach this remains an open question. For instance, one side argues about the morality of exposing participants to harm, while another points to the importance of internet safety. Exercising counterspeech can offer mitigation of online abuse in good faith and there are legal groundings that can potentially be applied to encourage such an action. As an example, **Good Samaritan laws** provide indemnity to people who assist others in danger (Smits, 2000). These safeguards aim to ensure that individuals are not hesitant to help others in distress due to the fear of facing legal consequences in case of unintentionally making errors in their efforts to provide support. In 2017 the EU Commission released a communication emphasizing the need to tackle illegal content online, stating that '*This Communication ... aims to provide clarifications to platforms on their liability when they take proactive steps to detect, remove or disable access to illegal content (the so-called "Good Samaritan" actions)*' (Commission, 2017). We argue that this statement can be extended to the scenario of applying counterspeech to online hate mitigation.

Responsible open-source research can facilitate reproducibility and transparency of science. Recently, reproducible research has been deemed critical in both social sciences (Stroebe et al., 2012; Derksen and Morawski, 2022) and computer science, and low replication success is found despite using materials provided in the original papers (Belz et al., 2023; Collaboration, 2015). To tackle this issue, a few initiatives for transparent research have been proposed, advocating researchers to state succinctly in papers how experiments are conducted (e.g., stimuli, mechanisms for data selection) and evaluated, including A 21 Word Solution (Simmons et al., 2012) and Open Science Framework.[5] Furthermore, practising data sharing encourages researchers to be responsible for fair and transparent experimental designs, and to avoid subtle selection biases that might affect substantive research questions under investigation (Dennis et al., 2019). At the same time, when handling sensitive or personal information, data sharing should adhere to research ethics and privacy standards (Dennis et al., 2019; de la Cueva and Méndez, 2022). For instance, in the case of hate speech, using synthetic examples or de-identification techniques is considered a good general practice for ensuring the safety of individuals (Kirk et al., 2022).

• ***Automating counterspeech generation*** There are several ethical challenges related to automating the task of counterspeech generation. First of all, there is the danger of dual-use: the same methodology could also be used to silence other voices.

Furthermore, effective and ethical counterspeech relies on the accuracy and robustness of detecting online hate speech: an innocent speaker may be publicly targeted and shamed if an utterance is falsely classified as hate speech – either directly or indirectly as in end-to-end response generation. For example, Google's Jigsaw API (Google Jigsaw, 2022), a widely used tool for detecting toxic language, makes predictions that are aligned with racist beliefs and biases—for example it is less likely to rate anti-Black language as toxic, but more likely to mark African American English as toxic (Sap et al., 2022). It is thus important to make sure that the underlying tool is not biased and well-calibrated to the likelihood that an utterance was indeed intended as hate speech. For example, the 'tone' of counterspeech could be used to reflect the model's confidence.

A related question is free speech: what counts as acceptable online behaviour, what sort of speech is deemed inappropriate, in which contexts, and should be targeted by counterspeech? A promising direction for answering this complex question is participatory design to empower the voices of those who are targeted (Birhane et al., 2022).

In sum, there is a trade-off between risks and benefits of counterspeech generation. Following the 'Good Samaritan' law: automating counterspeech provides timely help to victims in an emergency which is protected against prosecution (even if it goes wrong). Similar legislation is adopted by other countries, including the European Union, Australia and the UK. Under this interpretation, well-intentional counterspeech (by humans and machines) is better than doing nothing at all.

Best practices We provide best practices for developing successful intervention tools.

1. Bear in mind practical use cases and scenarios of hate-countering tools. A single intervention strategy is unlikely to diminish online harm and successful counterspeech interventions would benefit from personalisation. To design successful counterspeech tools, it is important to consider the purposes of counter messages (e.g., support victims and debunk stereotypes), the speakers (e.g., practitioners, authorities and high-profile people), recipients (e.g., ingroup/outgroup, political background and education level), the content (e.g., strategy, style, and tones), intensity (e.g., one message per week/month), and the communication medium (e.g., videos, text, and platforms).

2. Look beyond automated metrics and consider deployment settings for evaluating the performance of generation systems. Generation systems are generally evaluated on test sets in a controlled environment using accuracy-based metrics (e.g., ROUGE and BLEU) that cannot address social implications of a system. Drawn from social science studies, metrics assessing social impact (e.g., user engagement), behavioural change (e.g., measure abuse reduction in online discourse) and attitude change (e.g., through self-description questionnaires) can be considered. A good intervention system is expected to pertain long-lasting effects.

3. Be clear about the methodology employed in experiments, open-source experimental materials (e.g., stimuli, questionnaires and codebook), and describe the desirable criteria for evaluating counterspeech intervention. As standardised procedures are not yet established for the assessment of counterspeech interventions, examining the impact of interventions becomes difficult. A meaningful description of experimental design would therefore enhance reproducible research and help capture the limitation of existing research.

4. Establish interdisciplinary collaboration across areas such as counter-terrorism, political science, psychology and computer science. AI researchers can help guide policymakers and practitioners to, for instance, identify long-term interventions by performing large-scale data analysis using standardized procedures on representative and longitudinal samples. With expertise in theories of human behaviour change and experimental design, social science researchers can conduct qualitative

[5] https://osf.io/

evaluations of AI intervention tools in real-life scenarios to understand their social impact.

8 Conclusion

Online hate speech is a pressing global issue, prompting scientists and practitioners to examine potential solutions. Counterspeech, content that directly rebuts hateful content, is one promising avenue. While NLP researchers are already beginning to explore opportunities to automate the generation of counterspeech for the mitigation of hate at scale, research from the social sciences points to many nuances that need to be considered regarding the impact of counterspeech before this intervention is deployed. Taking an interdisciplinary approach, we have attempted to synthesize the growing body of work in the field. Through our analysis of extant work, we suggest that findings regarding the efficacy of counterspeech are highly dependent on several factors, including methodological ones such as study design and outcome measures, and features of counterspeech such as the speaker, target of hate, and strategy employed. While some work finds counterspeech to be effective in lowering further hate generation from the perpetrator and raising feelings of empowerment in bystanders and targets, others find that counterspeech can backfire and encourage more hate. To understand the advantages and disadvantages of counterspeech more deeply, we suggest that empirical research should focus on testing counterspeech interventions in real-world settings which are scalable, durable, reliable, and specific. Researchers should agree on key outcome variables of interest in order to understand the optimal social conditions for producing counterspeech at scale by automating its generation. We hope that this review helps make sense of the variety of types of counterspeech that have been studied to date and prompts future collaborations between social and computer scientists working to ameliorate the negative effects of online hate.

Acknowledgements

We thank Bertie Vidgen for the valuable feedback on the initial structure of this manuscript and Hannah Rose Kirk for her help with the collection of target literature.

Gavin Abercrombie and Verena Rieser were supported by the EPSRC projects ‘Equally Safe Online’ (EP/W025272/1 and EP/W025493/1) ‘Gender Bias in Conversational AI’ (EP/T023767/1). Yi-Ling Chung, Florence Enock and Jonathan Bright were supported by the Ecosystem Leadership Award under the EPSRC Grant EPX03870X1 and The Alan Turing Institute.

References

Adak, Sayantan, Souvic Chakraborty, Paramita Das, Mithun Das, Abhisek Dash, Rima Hazra, Binny Mathew, Punyajoy Saha, Soumya Sarkar, and Animesh Mukherjee. 2022. Mining the online infosphere: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(5):e1453.

Akazawa, Nami, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Distilling implied bias from hate speech for counter narrative selection. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 29–43, Prague, Czechia. Association for Computational Linguistics.

Alsagheer, Dana, Hadi Mansourifar, and Weidong Shi. 2022. Counter hate speech in social media: A survey. *arXiv preprint arXiv:2203.03584*.

Ashida, Mana and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Bartlett, Jamie and Alex Krasodomski-Jones. 2015. Counter-speech examining content that challenges extremism online. *DEMOS, October*.

Belz, Anya, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Benesch, Susan. 2014a. Countering dangerous speech: New ideas for genocide prevention. *Washington, DC: US Holocaust Memorial Museum*.

Benesch, Susan. 2014b. Defining and diminishing hate speech. *State of the world’s minorities and indigenous peoples*, 2014:18–25.

Benesch, Susan, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on Twitter: A field study. *Dangerous Speech Project*.

Berend, Gábor. 2022. Combating the curse of multilinguality in cross-lingual WSD by aligning sparse contextualized word representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2459–2471, Seattle, United States. Association for Computational Linguistics.

Bertoldi, Nicola, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.

Bilewicz, Michał, Patrycja Tempska, Gniewosz Leliwa, Maria Dowgiałło, Michalina Tańska, Rafał Urbaniak, and Michał Wroczyński. 2021. Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, 47(3):260–266.

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bonaldi, Helena, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 13–28, Prague, Czechia. Association for Computational Linguistics.

Bonaldi, Helena, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. *arXiv preprint arXiv:2211.03433*.

Buerger, Catherine. 2021a. Counterspeech: A literature review. *Available at SSRN 4066882*.

Buerger, Catherine. 2021b. #iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media + Society*, 7(4):20563051211063843.

Buerger, Catherine. 2022. Why they do it: Counterspeech theories of change. *Available at SSRN 4245211*.

Bélanger, Jocelyn J., Claudia F. Nisa, Birga M. Schumpe, Tsion Gurmu, Michael J. Williams, and Idhamsyah Eka Putra. 2020. Do counter-narratives reduce support for isis? yes, but not for their target audience. *Frontiers in Psychology*, 11.

Carthy, S. L. and K. M. Sarma. 2021. Countering terrorist narratives: Assessing the efficacy and mechanisms of change in counter-narrative strategies. *Terrorism and Political Violence*, 0(0):1–25.

Carthy, Sarah L, Colm B Doody, Katie Cox, Denis O'Hora, and Kiran M Sarma. 2020. Counter-narratives for the prevention of violent radicalisation: A systematic review of targeted interventions. *Campbell Systematic Reviews*, 16(3):e1106.

Cettolo, Mauro, Nicola Bertoldi, and Marcello Federico. 2014. The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179.

Chakravarthi, Bharathi Raja. 2022. Multilingual hope speech detection in english and dravidian languages. *International journal of data science and analytics*, 14(4):389—406.

Chaudhary, Mudit, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An nlp perspective. *arXiv preprint arXiv:2109.02941*.

Chung, Yi-Ling, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Chung, Yi-Ling, Serra S. Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021b. Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Chung, Yi-Ling, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Online.

Chung, Yi-Ling, Serra Sinem Tekiroğlu, and Marco Guerini. 2021c. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Citron, Danielle Keats and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435.

Collaboration, Open Science. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Commission, European. 2017. Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

de la Cueva, Javier and Eva Méndez. 2022. Open science and intellectual property rights. how can they better interact? state of the art and reflections. report of study. european commission.

Dennis, Simon, Paul Garrett, Hyungwook Yim, Jihun Hamm, Adam F Osth, Vishnu Sreekumar, and Ben Stone. 2019. Privacy versus open science. *Behavior research methods*, 51:1839–1848.

Derksen, Maarten and Jill Morawski. 2022. Kinds of replication: Examining the meanings of “conceptual replication” and “direct replication”. *Perspectives on Psychological Science*, 17(5):1490–1505. PMID: 35245130.

Dinan, Emily, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland. Association for Computational Linguistics.

Dušek, Ondřej and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Ernst, Julian, Josephine B Schmitt, Diana Rieger, Ann Kristin Beier, Peter Vorderer, Gary Bente, and Hans-Joachim Roth. 2017. Hate beneath the counter speech? A qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, (10):1–49.

Fanton, Margherita, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Ferguson, Kate. 2016. Countering violent extremism through media and communication strategies: A review of the evidence.

Finnegan, Eimear, Jane Oakhill, and Alan Garnham. 2015. Counter-stereotypical pictures as a strategy for overcoming spontaneous gender stereotypes. *Frontiers in Psychology*, 6.

Fortuna, Paula, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Frenkel, Sheera and Kate Conger. 2022. Hate Speech’s Rise on Twitter Is Unprecedented, Researchers Find. *The New York Times*.

Furman, Damián, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, Singapore. Association for Computational Linguistics.

Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Garland, Joshua, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1):3.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Goffredo, Pierpaolo, Valerio Basile, Bianca Cepollaro, and Viviana Patti. 2022. Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 57–66, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Google Jigsaw. 2022. Perspective API. Accessed: 26 May 2023.

Gupta, Rishabh, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.

Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Murias Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Hassan, Sabit and Malihe Alikhani. 2023. DisCGen: A framework for discourse-informed counterspeech generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 420–429, Nusa Dua, Bali. Association for Computational Linguistics.

He, Bing, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2022. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 90–94, New York, NY, USA. Association for Computing Machinery.

Iqbal, Khuram, Saad Kalim Zafar, and Zahid Mehmood. 2019. Critical evaluation of pakistan's counter-narrative efforts. *Journal of Policing, Intelligence and Counter Terrorism*, 14(2):147–163.

Jay, Timothy. 2009. Do offensive words harm people? *Psychology, public policy, and law*, 15(2):81.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Jiang, Shuyu, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.

Kennedy, Chris J, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Kirk, Hannah, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leader Maynard, Jonathan and Susan Benesch. 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention*, 9(3).

Lee, Huije, Young Ju NA, Hoyun Song, Jisu Shin, and Jong C. Park. 2022. Elf22: A context-based counter trolling dataset to combat internet trolls. In *Proceedings of the 13th Language Resources and Evaluation, LREC 2022, Marseille, France, June 20-25, 2022*, pages 3530–3541. European Language Resources Association.

Leonhard, Larissa, Christina Rueß, Magdalena Obermaier, and Carsten Reinemann. 2018. Perceiving threat and feeling responsible. how severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on facebook. *Studies in Communication and Media*, 7(4):555–579.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, Hao, Pradeep Nalluri, Lantian Li, Yifan Sun, and Yongjun Zhang. 2022. Multiplex anti-Asian sentiment before and during the pandemic: Introducing new datasets from Twitter mining. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 16–24, Dublin, Ireland. Association for Computational Linguistics.

Litaker, John R., Carlos Lopez Bray, Naomi Tamez, Wesley Durkalski, and Richard Taylor. 2022. Covid-19 vaccine acceptors, refusers, and the moveable middle: A qualitative study from central texas. *Vaccines*, 10(10).

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Mathew, Binny, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on Twitter. *arXiv:1812.02712*.

Mathew, Binny, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, 151(4):264–269. PMID: 19622511.

Möhle, Pauline, Matthias Orlikowski, and Philipp Cimiano. 2023. Just collect, don't filter: Noisy labels do not improve counterspeech collection for languages without annotated resources. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 44–61, Prague, Czechia. Association for Computational Linguistics.

Mun, Jimin, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore. Association for Computational Linguistics.

Munger, Kevin. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

News, BBC. 2018. MPs 'being advised to quit Twitter' to avoid online abuse. *BBC News*.

Nie, Feng, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy. Association for Computational Linguistics.

Novikova, Jekaterina, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Obermaier, Magdalena, Desirée Schmuck, and Muniba Saleem. 2021. I'll be there for you? effects of islamophobic online hate speech and counter speech on muslim in-group bystanders' intention to intervene. *New Media & Society*, 0(0):14614448211017527.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Poole, Elizabeth, Eva Haifa Giraud, and Ed de Quincey. 2021. Tactical interventions in online hate speech: The case of #stopislam. *New Media & Society*, 23(6):1415–1442.

Priyadharshini, Ruba, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Procter, Rob, Helena Webb, Marina Jirotka, Pete Burnap, William Housley, Adam Edwards, and Matt Williams. 2019. A study of cyber hate on twitter with implications for social media governance strategies. *arXiv preprint arXiv:1908.11732*.

Qian, Jing, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Reynolds, Louis and Henry Tuck. 2016. The counter-narrative monitoring & evaluation handbook. *Institute for Strategic Dialogue*.

Riedl, Martin J., Gina M. Masullo, and Kelsey N. Whipple. 2020. The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, 107:106262.

Saha, Koustuv, Eshwar Chandrasekharan, and Munmun De Choudhury. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 255–264, New York, NY, USA. Association for Computing Machinery.

Saha, Punyajoy, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech.

Saltman, Erin, Farshad Kooti, and Karly Vockery. 2021. New models for deploying counterspeech: Measuring behavioral change and sentiment analysis. *Studies in Conflict & Terrorism*, 0(0):1–24.

Saltman, Erin Marie and Jonathan Russell. 2014. White paper–the role of Prevent in countering online extremism. *Quilliam publication*.

Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Schieb, Carla and Mike Preuss. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ICA annual conference, at Fukuoka, Japan*, pages 1–23.

Siegel, Alexandra A. 2020. Online hate speech. *Social media and democracy: The state of the field, prospects for reform*, pages 56–88.

Silverman, Tanya, Christopher J Stewart, Jonathan Birdwell, and Zahed Amanullah. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue*.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2012. A 21 word solution. *Available at SSRN 2160588*.

Smits, Jan M. 2000. The good samaritan in european private law; on the perils of principles without a programme and a programme for the future.

Snyder, C. R., Kevin L. Rand, and David R. Sigmon. 2018. Hope Theory: A Member of the Positive Psychology Family. In *The Oxford Handbook of Hope*, pages 257–276. Oxford University Press.

Solaiman, Irene, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Stroebe, Wolfgang. 2008. Strategies of attitude and behaviour change.

Stroebe, Wolfgang, Tom Postmes, and Russell Spears. 2012. Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7(6):670–688.

Stroud, Scott R and William Cox. 2018. The varieties of feminist counterspeech in the misogynistic online world. *Mediating Misogyny: Gender, Technology, and Harassment*, pages 293–310.

Tekiroğlu, Serra Sinem, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Tekiroğlu, Serra Sinem, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Toliyat, Amir, Sarah Ita Levitan, Zheng Peng, and Ronak Etemadpour. 2022. Asian hate speech detection on twitter during covid-19. *Frontiers in Artificial Intelligence*, 5.

Tuck, Henry and Tanya Silverman. 2016. *The counter-narrative handbook*. Institute for Strategic Dialogue.

Vallecillo-Rodríguez, Maria Estrella, Arturo Montejo-Raéz, and Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71:227–245.

Vidgen, Bertie, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Vidgen, Bertie, Helen Margetts, and Alex Harris. 2019. How much online abuse is there? A systematic review of evidence for the UK. *Alan Turing Institute Policy Briefing*.

Vidgen, Bertie, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Wang, Ke and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.

Wright, Lucas, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62.

Yin, Wenjie and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Yu, Xinchen, Eduardo Blanco, and Lingzi Hong. 2022. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States. Association for Computational Linguistics.

Yu, Xinchen, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289, Singapore. Association for Computational Linguistics.

Zellers, Rowan, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zheng, Yi, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

Zhou, Chunting, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Zhu, Wanzheng and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Documenting Geographically and Contextually Diverse Language Data Sources

Angelina McMillan-Major[1]*, Francesco De Toni[2]*, Zaid Alyafeai[3], Stella Biderman[4,5] Kimbo Chen[6,7], Gérard Dupont[8], Hady Elsahar[9,10], Chris Emezue[9,11], Alham Fikri Aji[12], Suzana Ilić[13], Nurulaqilla Khamis[14], Colin Leong[9,15], Maraim Masoud[7], Aitor Soroa[16], Pedro Ortiz Suarez[17], Daniel van Strien[18], Zeerak Talat[12]*, Yacine Jernite[18]

University of Washington[1], Australian National University[2], ARBML[3], Booz Allen Hamilton[4], EleutherAI[5], BigScience[6], Independent Researcher[7], Mavenoid[8], Masakhane[9], Meta FAIR[10], Mila[11], Mohamed Bin Zayed University of Artificial Intelligence[12], University of Innsbruck[13], Faculty of Electrical Engineering, Universiti Teknologi Malaysia[14], University of Dayton[15], University of the Basque Country[16], Common Crawl Foundation[17], Hugging Face[18]

aymm@uw.edu, francesco.detoni@anu.edu.au, z@zeerak.org

Abstract Contemporary large-scale data collection efforts have prioritized the amount of data collected to improve large language models (LLM). This quantitative approach has resulted in concerns for the rights of data subjects represented in data collections. This concern is exacerbated by a lack of documentation and analysis tools, making it difficult to interrogate these collections. Mindful of these pitfalls, we present a methodology for documentation-first, human-centered data collection. We apply this approach in an effort to train a multilingual LLM. We identify a geographically diverse set of target language groups (Arabic varieties, Basque, Chinese varieties, Catalan, English, French, Indic languages, Indonesian, Niger-Congo languages, Portuguese, Spanish, and Vietnamese, as well as programming languages) for which to collect metadata on potential data sources. We structure this effort by developing an online catalogue in English as a tool for gathering metadata through public hackathons. We present our tool and analyses of the resulting resource metadata, including distributions over languages, regions, and resource types, and discuss our lessons learned.

1 Introduction

Current trends in developing large language models (LLM) require the use of vast amounts of data (Brown et al., 2020; Gao et al., 2020; Rae et al., 2021). Typically, this data is collected from online sources, ranging from highly edited and structured text such as Wikipedia to the myriad text and audiovisual components of web pages, e.g., collected by the Common Crawl Foundation.[1] However, recent research has raised concerns about the creation and use of such data resources. For instance, Wikipedia is highly biased in terms of the topics covered and the demographics of its contributors, particularly along gender, race, and geographic lines (Barera, 2020), resulting in concerns of representation in the technologies developed on Wikipedia-derived data. Data from Common Crawl has similarly been shown to correlate with country-level population density, relative access to the internet, and per capita GDP (Dunn, 2020) and to contain significant amounts of hate speech and sexually explicit content (Luccioni and Viviano, 2021). Irrespective of the data source, typical web-crawling collection practices have no structures for supporting informed consent beyond websites' own policies that users rarely read (Cakebread, 2017; Obar and Oeldorf-Hirsch, 2020).

Several documentation schemas for natural language processing (NLP) datasets (Bender and Friedman, 2018; Gebru et al., 2021; Holland et al., 2018; Stoyanovich and Howe, 2019; McMillan-Major et al., 2023) have been proposed to aid NLP researchers in documenting their datasets (Gao et al., 2020; Biderman et al., 2022; Gehrmann et al., 2021; Wang et al., 2021) and to retrospectively document and analyze datasets that were developed and released by others without thorough documentation (Bandy and Vincent, 2021; Kreutzer et al., 2022; Birhane et al., 2021; Dodge et al., 2021). Data docu-

[*] Corresponding Authors: Angelina McMillan-Major, Francesco De Toni, Zeerak Talat.

[1] http://commoncrawl.org/

Licensed under CC BY 4.0, pages 50–77, DOI 10.3384/nejlt.2000-1533.2024.5217

mentation to support transparency has gained traction, following calls for a reevaluation of the acquisition and use of data in machine learning (ML) at large (Birhane and Prabhu, 2021; Jo and Gebru, 2020; Paullada et al., 2021; Gebru et al., 2021; Bender et al., 2021). Building on this work, we propose a documentation-first and human-centered method for data collection for NLP that emphasizes consent, representation, self-determination, and privacy. Using this method, we create a data catalogue for training multilingual LLMs that promotes responsible data collection and data subjects' rights to control over their own data. We conclude that starting documentation processes during the data collection phase can contribute to building a more representative dataset and allows for early identification of ethical concerns. Our contributions consist of the data catalogue tool,[2] which remains openly available for use in collecting metadata and for searching existing entries, as well as the human-centered methodology of data collection in collaboration with language communities for representative language modeling and other NLP tasks.

1.1 Research Context

Our work was situated within a large-scale global coalition of experts in NLP and related fields dedicated to researching questions related to language modeling known as the BigScience Workshop.[3] The BigScience Workshop was started as an open collaboration of international researchers by Hugging Face, GENCI (Grand Equipement National de Calcul Intensif), and IDRIS (The Institute for Development and Resources in Intensive Scientific Computing) and was dedicated to open research of NLP, social sciences, and the legal, ethics and public policy of large language models. While this coalition (henceforth *the workshop*) had many working groups with different foci determined by the research interests of the participating researchers, one of its primary goals was to train and publicly release a multilingual LLM. Key to this endeavor was the creation of a dataset to train the model on.

Bearing in mind the limitations of prior large-scale data collection efforts, we aimed to intentionally curate our dataset for *representativeness*. We defined representativeness based on the intersection of geographic and sociolinguistic contexts. This means that, for each target language, we aimed to collect data for the relevant dialects and regions where that language is spoken. Like most language modeling endeavors, we relied on commonly used web sources for collection, but we also highlighted the need for other formats, including books, audio from radio programs and podcasts, and others. Starting from this goal and the coalition members' languages of expertise, we identified 13 language groups to target for inclusion in the model training: Arabic varieties, Basque, Chinese varieties, Catalan, English, French, Indic languages, Indonesian, Niger-Congo languages, Portuguese, Spanish, and Vietnamese, as well as programming languages. In addition to coalition members speaking many of these languages themselves, we were also motivated to intentionally select data resources for these languages in order to improve the resulting language model's performance in generating these languages. Programming languages were included in the design of the language model, but because they are not natural languages with communities of use, we did not organize a specific hackathon to collect entries for them in the catalogue (see §5).

1.2 Overview

We prepared for the challenges of responsible dataset creation by focusing our efforts on documenting potential sources prior to their collection. Meanwhile, other working groups on data governance and data tooling created pipelines for hosting and processing data. In the next sections, we compare our documentation effort (henceforth *the catalogue*) to already developed catalogs in linguistics and NLP (§2). In §3 we present our catalogue and associated online form[4], including our process for designing the catalogue.

We developed the online submission form to facilitate public hackathon events for collecting metadata for language resources from specific regions (§4). While the form prioritizes submitting entries for the target languages, we made it possible for entries for any language to be submitted as the catalogue remains open for submissions and browsing after the end of the hackathons. Although the catalogue is a living documentation effort, we present the results obtained after the initial documentation effort (§5). We then discuss lessons learned in creating the catalogue, its potential use as a model for data documentation endeavors in NLP, and the limitations of our approach, suggesting improvements for future data documentation efforts (§7). Finally, we consider the ethical implications of our approach, especially with regard to data licensing and personally identifiable information (§8).

2 Related Work

Since the early 90s, NLP data organizations have maintained catalogs for datasets and tools in order to support language research.[5] While the metadata for these

[2] Available at `https://bigscience.huggingface.co/data-catalogue-form`

[3] `https://bigscience.huggingface.co/`

[4] See Footnote 2 for URL.

[5] Organizations include the Linguistic Data Consortium (LDC), The Southern African Centre for Digital Language Resources (SADiLaR),

catalogs are openly available, accessing the language resources (e.g., annotated corpora and lexicons) and supporting tools may require paying for a license to the resource or for membership to the catalog. The fees support the creation, licensing, storage, and maintenance of new datasets and language research initiatives. The LDC, for example, currently provides access to 1016 datasets.[6]

Open source dataset catalogs have also been constructed as supporting technical infrastructure in the context of NLP and ML libraries. The Natural Language Toolkit (NLTK), developed since 2001, is a Python package with utilities for NLP tasks that includes access to widely used corpora such as the Brown Corpus (Kučera and Francis, 1967) as well as features for adding datasets and using datasets locally (Bird et al., 2009). The Hugging Face Datasets library (Lhoest et al., 2021) and Tensorflow library (Tensor Flow Authors, 2021) both provide tools for loading datasets from remote and local repositories and include catalogs of directly accessible datasets. SADiLaR provides its own catalog of annotated language datasets and processing tools, with links for downloading resources that are licensed for distribution. Other catalogs of NLP datasets do not provide access to the datasets themselves, but provide information about uses and categories. For example, Papers with Code links academic publications that use the same dataset with information about the dataset.[7] Masader similarly provides metadata about Arabic-language NLP datasets without hosting the data (Alyafeai et al., 2022).

Our work is an effort to merge the careful and well-established data collection and documentation practices from organizations such as the LDC with the collaborative, open source tools for dataset construction. While large-scale NLP research requires vasts amounts of data, the work that goes into curating, documenting, and maintaining the data is often undervalued (Sambasivan et al., 2021), resulting in data collections that are often too large to document post-hoc (Bender et al., 2021) and contain significant quantities of unwanted media (Luccioni and Viviano, 2021). We provide an alternate approach to data collection and management in NLP; this approach prioritises documentation in the data creation process, engages communities to inform data curation, and contributes to a more representative dataset.

3 The Catalogue

The primary goal of the catalogue (see appendix A for screenshots of the form) was to support the creation of a training dataset for language modeling that integrated with the efforts of the other working groups and aligned with the values defined by the workshop governance. We surveyed each working group to identify their particular metadata needs, resulting in almost 40 categories of metadata. Aiming to balance the information needs of the working groups with the effort required to submit a resource and its metadata to the catalogue, we grouped and prioritized the categories. We further prioritized metadata that are applicable across as many languages and data sources as possible. We did not make use of existing metadata formalisms as we expected that they would discourage submissions to the catalogue by those unfamiliar with them. Instead we envisioned our metadata collection as an upstream process that would be flexible enough to contribute to many different kinds of downstream annotation or metadata labeling tasks.

We created an openly accessible form in English for submitting metadata for potential sources for the identified language groups.[8] We used an iterative approach to collectively develop questions that elicit the metadata, descriptions of the information being requested, and answer prompts to support efficient documenting. Wherever possible, we formatted the questions as multiple choice questions with an optional free-form field, should the pre-existing options be insufficient. After building the online form, we tested the form with actual examples, i.e., the *Le Monde* newspaper and its publishing company *Group Le Monde* to ensure its validity.

3.1 The Catalogue Submission Form

Testing the form using the *Le Monde* newspaper example helped us update our form by surfacing discrepancies in specific questions for certain resource types, particularly concerning data processing. With this consideration in mind, we defined the following resource types: **primary source**, a single source of language data (text or speech), such as a newspaper, radio, website, or book collection; **processed language dataset**, a processed NLP dataset containing language data that can be used for language modeling; and **language organization or advocate,** an organization or person holding or managing language sources of various types, formats, and languages. We follow Jernite et al. (2022) in distinguishing between **data subjects** (those talked to or about in the data), **data creators** (those who create the text, audio, or video data), and **data custodians** (those who own or manage the data). We distinguish between a data custodian, who is responsible for handling requests for the data, and language organizations, that may ultimately hold the rights to the data but do not handle day-to-day requests, though in many cases the data custodian and the language organization of a resource are the same entity.

the European Language Resource Association (ELRA), the Chinese LDC, the LDC for Indian languages (LDCIL), and CLARIN.

[6] LCD Catalog by Year, accessed April 18, 2023.

[7] `https://paperswithcode.com/datasets`

[8] We built the form using Streamlit.

For all resource types, the form requests information about the languages and locations of the resource's data creators as well as contact information for a representative, owner, or custodian of the resource. Further questions are added for primary sources and processed datasets, including the availability of the resource and legal considerations for using the data, such as licenses, the type of data it contains, and the medium of the data.

3.1.1 General Information

The form first requests the source type, and then updates the questions once a type is selected. The following questions in the section request general information (e.g., a resource name, a unique identifier for searchability, and the resource's webpage). The form provides space for a description to display when searching the catalogue.

3.1.2 Languages and Locations

We designed the *Languages and Locations* section to accommodate various degrees of granularity in order to support and evaluate our goal of representativeness, and maximize the usability of the catalogue beyond the consortium's immediate use-case. The authors of each entry can specify what languages are represented in the resource by choosing from drop-down lists of our target language groups, with additional sub-lists for languages in the Indic and Niger-Congo families, and other languages as defined by the BCP-47 standard (Phillips and Davis, 2009). The form also provides space for submitting comments about the language variety in the resource, such as whether it contains language data that exhibits dialectal variation or code-switching. Similarly, authors can add information about the geographical origin of the data (i.e., the primary location of the language creators whose data is captured in the resource) using a drop-down list of macroareas ranging from world-wide to continents to regions (such as Western Africa or Polynesia) in addition to specific countries, nations, regions, and territories.

3.1.3 Representative, Owner, or Custodian

Responsible dataset creation includes respecting the rights of the data custodian, the person or organization that owns or manages the data source. The form allows for linking the resource being submitted to an existing organization in the catalogue via a drop-down list. If the data custodian is not already in the catalogue as a language organization, the remaining questions elicit their name, type, location, and contact information. This information supports our own and future catalogue users' efforts to understand local legal structures, communicate with data custodians about data use, and request permission for uses beyond those granted by licenses.

3.1.4 Availability of the Resource

For primary sources and existing datasets, the form requests information about how to obtain the data, i.e., through a link or contacting the data custodian. Depending on the response, the form asks for the URL to download the data or the data custodian's contact information. In characterizing the licenses or terms of use, the form asks whether the resource is accompanied by an explicit license. If the license or terms are known, the submitter may select a description such as public domain, research use, non-commercial use, or do not distribute. Submitters can also select relevant licenses from a drop-down list of frequently used licenses, or input the terms or license text into the form. If the licensing terms are unknown or unclear, the form requests that the submitter gives their best assessment of whether the data can be used to train models while respecting the rights and wishes of the data subjects, creators, and custodians.

3.1.5 Primary Source Type

The form allows for characterizations of the resource data for both primary sources and processed language datasets. We provide options for two kinds of resource descriptions. **Collections** may contain books or publishers, scientific articles and journals, news articles, radio programs, movies and documentaries, podcasts, or a user-suggested response. **Websites** may include social media, forums, news or magazine websites, wikis, blogs, content repositories, or a user-suggested response.

If the submission is a processed language dataset, the section appears in the form as *Primary Sources of the Processed Dataset*. If the dataset contains original data, no further questions appear. If the data is a collection of primary sources, the form presents questions about those sources, such as if they are openly available or have accessible documentation. Users may link the processed dataset to primary sources already documented in the catalogue or provide original descriptions of those primary sources. The final question concerns the licensing information of the primary sources, as these may differ from the dataset itself. See §8.1 for further discussion.

3.1.6 Media Type, Format, Size, and Processing

The final section of the form addresses the technical aspects of the resource. A submitter may indicate the medium of the data (text, audiovisual data, images, or a combination thereof) and details about the data format (the file type or distribution format). If the data includes text, the form asks if the text was transcribed. While most datasets appear with metadata about the size of the data given by mega- or gigabytes, primary sources often do not have this information available. Instead, we asked submitters to provide an estimate of the amount of data using a descriptive unit of data, i.e., articles, posts, episodes, books, webpages, or a user-provided unit. The form then asks for the number of instances in the resource using the provided unit and the average number of words in the unit using ranges of magnitudes of 10. This information was useful to the coalition's data processing working groups, but it proved difficult for the submitters to estimate, unless already available in the source metadata. On completion, submitters could review their responses as it will be saved (in a JSON format) before submitting their entry to the catalogue.

4 Additional Features

There are two other modes for interacting with the catalogue: a validation mode, for validating submitted entries, and a visualization mode, for filtering and mapping specific submitted entries. Because we intended to make the catalogue openly available on the web past the end date of the workshop, we included the validation functionality to allow users to confirm that metadata for submitted entries was correct and could be updated if ever the information was no longer correct (e.g., if a license for an entry changed). The purpose of the visualization mode was to support later users of the catalogue in seeing the general distribution of submitted resources of the catalogue across languages and geographic regions and searching for specific resources within those categories.

To validate an entry, the validator can confirm the previously submitted metadata or edit and resubmit the entry. The catalogue then saves both the original and the validated submission. The visualizations include a pie chart detailing the proportion of entries by language and an interactive map which shows the number of submitted entries for a region or country as defined by the location of the data creators or data custodians. In Figure 1, the color gradient indicates the number of entries by country and location markers indicate regions that can be examined for more details. Both the map and a pie chart can be filtered using one of the many properties produced by the form, e.g., the resource, license, or media type. Entries returned by the filter can be selected to display their descriptions.

5 Community Hackathons

With the catalogue submission form developed, we could begin to collect and document potential data sources for review prior to developing the full dataset towards the workshop's LLM goal. Whereas prior data collection processes utilized automatic methods for collecting as much data as possible, we wanted our collection process to prioritize sources that were created by language communities and that were determined by language communities to be representative of their language use. In order to center the metadata collection for as many languages as possible around communities who speak those languages, we decided to crowdsource our metadata collection by organizing community hackathons.[9] To do so, we reached out to regional community organizations focused on ML and NLP to collaborate in leading local hackathons and put out a similar call within the workshop for individuals who spoke one or more of the listed languages. The task for each hackathon was for participants to use the catalogue submission form to submit as much metadata as they could find on potential data sources for their language or languages. We developed a guide[10] with instructions and suggestions for the hackathon participants for each section of the catalogue submission form. A coalition member and/or a collaborating organizer from a partner organization was available to interact with participants and answer questions arising while filling the form and to discuss details about potential resources or institutions.

In total, we organized 6 hackathons for specific communities and regions of the world based on the availability of organizers and their familiarity with the communities, namely African languages in collaboration with Masakhane,[11] Asian languages with Machine Learning Tokyo,[12] Basque, English in the Americas and Europe, English in the Indo-Pacific Region, and Spanish in Latin America with LatinX in AI.[13] The hackathons took place online in October-December, 2021, lasting one to six hours. We announced hackathons using social media, in coordination with the relevant partner organizations. Because we advertised primarily to members of the workshop, social media followers of the workshop, and members of the partner organizations, the hackathons attracted participants who were generally interested in language modeling and specifically wanted to support the workshop goals of having greater language representation in the to-be-trained workshop language model. No further incentives were used to encourage participation. During the hackathons we only collected a name and e-mail. After the hackathons, we sent a 10-question survey to all participants to collect further information.

[9] Because programming languages are not natural languages with communities of speakers or signers, we did not organize a hackathon focused on programmming languages.

[10] Available at https://github.com/bigscience-workshop/data_sourcing/blob/master/sourcing_sprint/guide.md.

[11] https://www.masakhane.io/

[12] https://www.mlt.ai/

[13] https://www.latinxinai.org/

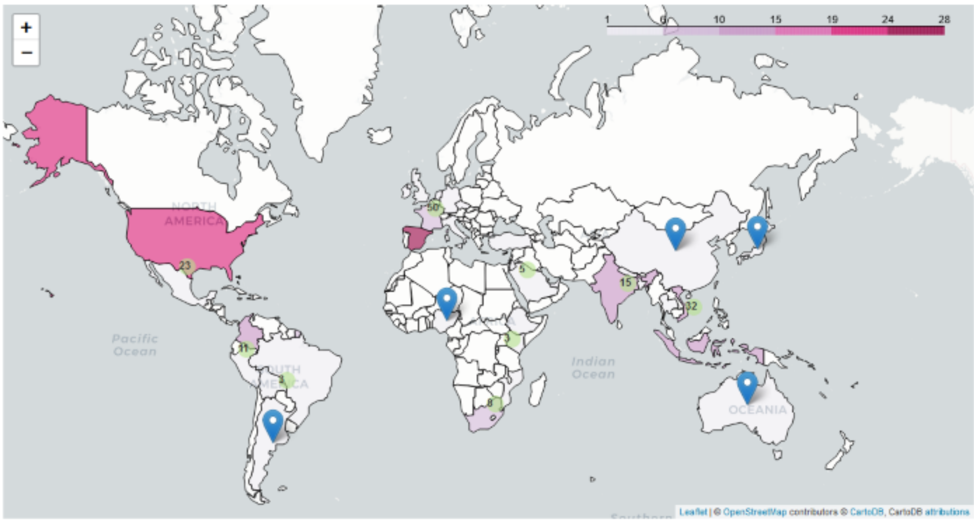


Figure 1: Geographical visualization of the locations of entries' data custodians. The color gradient indicates the number of entries by country and location markers indicate regions that can be examined for more entries and details.

6 Results

6.1 Hackathon Participation

Forty-one participants submitted descriptions of resources to the catalogue during the hackathons, of whom 11 responded to the survey. The first survey questions focused on participants' professional context, i.e., the country they are located in, their field of study and current stage in their career. The respondents were from diverse geographical location and career stages. Four respondents were located in Spain, with 3 in the Basque Country, while the remaining respondents were located in France, Japan, Kenya, Singapore, Sweden, Taiwan, and the USA. Respondents' career stages ranged from undergraduate student to a senior level position in industry, though most (7) listed an academic position. The most common research interests were NLP (8), data science (5), and linguistics (4). Other interests included library and/or information science, ethics/safety, recommendation systems, vision, creative AI, and optimization and compression techniques.

The remaining questions concerned participants' experiences before and during the hackathons. Most participants became aware of the hackathons through the coalition's internal channels or the communities and organizations that collaborate with us. Only two respondents listed social media as their entry point. Most respondents (6) only submitted resources for languages that they were fluent or advanced speakers of, while three respondents contributed resources that covered almost all of the target languages, most of which they had no familiarity with. In describing their motivations for participating in the hackathons, the most common reasons included developing the training dataset, supporting under-resourced languages in general, and improving the coverage of a particular language.

6.2 Gathered Resources

After the sixth and final hackathon, the catalogue contained 192 entries with 955 different language tags.[14] The most frequent language tags were those of the target language groups. Figure 2 shows the distribution of the target language groups across entries.[15] English is the most frequent language across all entries. For Arabic, the most frequent varieties are Modern Standard Arabic (13) and Classical Arabic (5). All other variants have 2 or fewer entries. The most frequent Indic languages are Hindi (15), Bengali (11), Telugu (9), Tamil (9), and Urdu (8) and the most frequent Niger-Congo languages are Swahili (9), Igbo (7), Yoruba (6), and isiZulu (4), with other languages having no more than 3 entries.

On the other end of the spectrum, 380 languages were tagged only in 1 or 2 entries. However, some of these languages belong to broader target language groups: i.e., 10 languages from the Niger-Congo group (Sesotho, Kirundi, Bambara, Kinyarwanda, Chi Chewa, Wolof, Twi, Lingala, ChiShona, and Kikuyu), and 12 varieties of Arabic (Algeria, Djibouti, Gulf, Egypt, Levant, Libya, Mauritania, Morocco, North Africa, Somalia, South Sudan, Sudan). Digitally accessible resources for these language varieties are less common than digital resources for languages with more frequent use on the internet, in part due to the smaller sizes of the com-

[14] The list of language tags includes both Arabic (generic tag) and specific varieties of Arabic (e.g. Classical Arabic). The form remains open and new entries have been added since the final hackathon. At present, the there are 252 entries in the catalogue.

[15] Due to multilingual resources, the percentages exceed 100%.

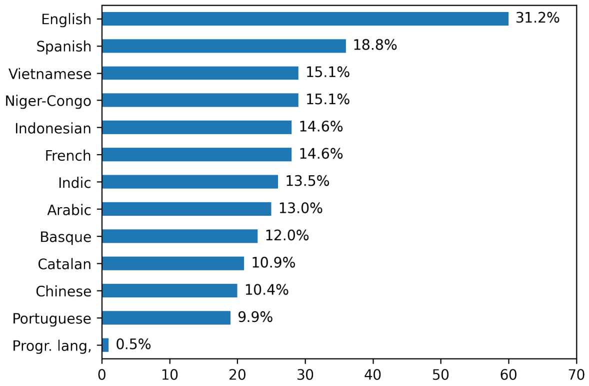


Figure 2: Relative distribution of the target languages in absolute values and as percentages of the total number of entries.

Language location	#	Percentage of all entries
Africa	18	9.38%
Americas*	3	1.56%
Asia	61	31.77%
Europe	46	23.96%
Latin America and the Caribbean	17	8.85%
Middle East and North Africa	4	2.08%
North Africa	2	1.04%
North America	11	5.73%
Oceania	5	2.60%
World-wide	21	10.94%

* entries not specifying if N. Am. or Lat. Am. and the Car.

Table 1: Distribution of language locations according to data creators (not custodians) over geographic regions (only first location for each entry).

Location	Languages			
	En.	Fr.	Sp.	Port.
Africa*	6	4	0	1
Americas†	0	1	2	1
Asia	10	0	0	1
Europe	13	13	11	5
Latin America and the Carib.	3	0	15	2
North America	13	1	2	1
Oceania	5	0	0	0
World-wide	16	11	10	11

* including entries from North Africa; no entries from Middle East were recorded for these languages
† entries not specifying if N. Am. or Lat. Am. and the Car.

Table 2: Distribution of entries in English, French, Spanish and Portuguese across continents.

munities using these languages and in part due to the numerous sociopolitical factors that have led to the valuation and resource allocation towards some languages (usually associated with colonial powers) over others. Excluding these, 358 languages were tagged only once or twice.

The submissions to the catalogue show a clear bias towards certain languages: English and Spanish submissions accounted for about half of the target languages recorded by the end of the hackathons. On the other hand, Chinese is included in fewer entries than languages that have fewer speakers, e.g., French, Spanish and Vietnamese (see Eberhard and Fennig 2021). This imbalance is the result of the varying availability of sources across different languages and the linguistic expertise of the coalition and hackathon participants.

We did not require users to adhere to a strict taxonomy of geographic location (e.g., continent → country → region) when providing geographic locations of a source. The submitters could freely label their submissions by macroscopic area (e.g., a continent or macroregion within a continent), country, region within a country or some combination of these. These labels are then saved in a list of location tags for each entry. We made this design decision to simplify the process of selecting geographic location for submitters while avoiding nested questions with increasing geographic granularity, providing flexibility in geographic labelling. For example, it may make more sense to label resources in Arabic as from *Middle East and Northern Africa*, rather than from *Africa* and *Asia*, even though *Middle East and Northern Africa* does not denote a continent in geographic terms. As a result, the catalogue does not conform to a particular taxonomy but can provide a frequency distribution over the location tags.

We focus our analysis of the geographic distribution of the recorded languages on continents and macroregions (i.e., usually the first geographic area provided). For the small number of cases where only a country was provided, we manually assigned the information to their respective continent or macroregion. We see in Table 1 that more than half of the primary language locations of the entries are located in Asia and Europe.

We further manually grouped locations into continents and macroregions and investigated how regional varieties of English, French, Spanish and Portuguese entries are represented (see Table 2). We see that these languages are well represented in their European varieties. However, each language also has a number of entries from other geographical areas, which are language specific, and several entries that were tagged as 'World-wide' (entries that include examples of a target language from multiple geographies or multilingual sources).

Primary sources were the most common source type entered. Of the 192 entries, 98 (51%) are primary sources, 64 (33%) are processed datasets, and 30 (16%) are organizations (see Table 3 for distributions of source types across target language groups). With the exception of Catalan, Indic and Vietnamese, the target language groups have more primary sources than secondary sources.

Languages	Types		
	Primary	Processed	Org.
Arabic	13	3	9
Basque	15	0	8
Catalan	1	14	6
Chinese	9	4	7
English	29	13	18
French	13	4	11
Indic	8	11	7
Indonesian	15	8	5
Niger-Congo	11	5	13
Portuguese	7	3	9
Programming	1	0	0
Spanish	17	2	17
Vietnamese	8	15	6

Table 3: Distribution of the target languages in the catalogue across source types.

The largest share of sources recorded are stewarded by non-commercial entities (see Table 4). University and research institutions are the most frequent custodian type (23.44%), followed by commercial entities (21.35%) and nonprofit entities/NGOs (13.5%). Twenty-four (12.5%) records do not specify a custodian.

Custodian type	#
University or research institution	45
Commercial entity	41
Nonprofit / NGO	26
Not Specified	24
Private individual	20
Government organization	17
Library, museum or archival institute	16
Community (incl. online)	2
Startup	1

Table 4: Distribution of custodian types.

In terms of the custodians' geographic diversity, 28 catalouge entries do not record a custodian location while the remaining 164 do. While the custodian locations reflect the diversity of the catalogue, they also show that an outsized share are located in the USA and European countries (see Table 5). All the other locations were only recorded once (Bolivia, Burundi, Czech Republic, Ethiopia, Hong Kong, Ireland, Italy, Kenya, Luxembourg, Mexico, Netherlands, Peru, Saudi Arabia, Scotland, Thailand, Turkey, United Arab Emirates) or twice (Argentina, Bangladesh, Brazil, Japan, Jordan, Mozambique, Nepal, Nigeria, Taiwan).

The license metadata suggests that the hackathon participants made efforts to submit sources with open licenses or without copyright (see Table 6).[16] Public domain or open license account for 37% of entries and another 37% are entered as not having licenses.

Custodian location	#	Custodian Location	#
Spain	27	France	9
USA	22	South Africa	6
Vietnam	14	UK	5
Indonesia	14	Australia	4
India	11	Germany	4
Colombia	10	China	3

Table 5: Top 12 most frequent custodian locations.

Licensing properties	#	Percentage of all entries
Missing	71	37%
Open license	56	29%
Copyright	30	16%
Non-commercial use	18	9%
Public domain	18	9%
Research use	10	5%
Multiple licenses	7	4%
Do not distribute	2	1%

Table 6: Distribution of licensing properties.

The hackathon submission entries contained primarily text data, as shown in Table 7. Two thirds of the resources contain only text data, while 5% contained text and image data, 4% contained text and audiovisual data, and another 6% contained text, image, and audiovisual data combined. Only 3% of the resources contained solely audiovisual data. A further 16% of the resources are missing information about the media types contained within the resource. This may suggest that the resources were not accessible or did not provide sufficient documentation for the hackathon participants to determine the resource media types.

Media type	#	Percentage of all entries
Text only	128	66%
Text and image	9	5%
Text and audiovisual	7	4%
Text, image and audiovisual	11	6%
Audiovisual	5	3%
Missing	32	16%

Table 7: Distribution of media types.

After the hackathons, the resources within the catalogue data were downloaded and used to develop the

[16] Entries may have multiple license properties.

BigScience ROOTS Corpus. Details on the data processing methods for the dataset and the resulting data metrics may be found in Laurençon et al. (2022). While the resources from the catalogue were ultimately used in collaboration with other data sources such as the OSCAR version 21.09 corpus (Ortiz Suárez et al., 2019) in order to meet the data quota for training an LLM, the catalogue could continue to grow to provide more metadata on resources used for NLP tasks and support documentation efforts for future data collection projects.

7 Discussion

The result of our efforts is an openly available catalogue of 192 data sources, with each of our target natural language groups constituting at least 10% of the total submitted entries.[17] The majority of these resources are primary and processed resources, with data custodian primarily located in the Americas, Europe and Australia. The sources recorded in the catalogue were used as a core component of the training dataset for training the LLM developed by the coalition. In addition, the development of the catalogue serves as an opportunity for methodological reflection on documentation-first and human-centered data collection in NLP. In this section we discuss lessons learned from creating and crowdsourcing the catalogue and present recommendations for future data collection efforts.

7.1 Centering the Human

A human-centered approach to data is one that is focused on "human values such as privacy, human rights, and ethics", is engaged in "asking ... what [technology] should do", and is committed to "acknowledging and addressing the individuals, organizations, and communities behind ... data" (Shah et al., 2021, p. 794). Our collection methodology consists of engaging with language communities to prioritize the collection of resources that those communities deem are representative of their language, as opposed to automatic collection and language identification methods. Additionally, the form dedicates multiple sections to the individuals and groups that produced and hold the data while the hackathons made the data curation process more accessible to members of the coalition not working in the data-sourcing working group. Our methodology also centers humans by collecting information on the rights of data holders and owners (Jernite et al., 2022) prior to collecting the actual data. This affords making informed decisions with respect to privacy and ethical considerations as well as data curation choices for content. However, the methodology has several limitations. First, it only addresses the needs of immediate users of the catalogue. Serving less immediate data consumers and connecting them with data producers would require additional infrastructure. Second, the methodology does not protect the rights of data holders and owners from future malicious users of the catalogue, as it does not embed a data governance structure within it. We discuss these risks further in Section 8.

7.2 Representativeness

Representativeness across languages Our catalogue only covers a fraction of world languages, largely reflecting the languages and contexts of the coalition members; missing are signed languages, some of the most widely spoken languages, and most underrepresented languages. Moreover, the distribution of target language entries in the catalogue is not uniform. While the efforts of the hackathon resulted in diverse resources for the languages covered, especially for English, French, Spanish, and Portuguese, the variation in success across languages emphasizes the need to actively include collaborators who sign and speak underrepresented languages and supporting them in leadership positions in NLP research.

As our results evidence, our definition for success, namely broad geographical representation, had direct impacts on our ability to evaluate the catalogue. For instance, the loosely structured ontology for recording dialects and geographical locations on one hand provided users with flexibility to adapt data entry to each source. On the other, it becomes more difficult to analyze information across the catalogue. Aggregating dialects and geographical locations of data posed a challenge because sources may include examples from multiple dialects and/or regions, resulting in significant difficulties in creating a classification protocol that was applicable to all sources. Furthermore, information about the geographical location of the languages in the sources may not be easily accessible or available to submitters.

Representativeness beyond language diversity Our effort focused on ensuring geographic variation and representativeness of target languages. However, from the perspective of linguistics, representativness encompasses a broad set of variables. For example, Biber (1993) identifies 8 hierarchical parameters that define the representativeness of a corpus: the primary channel (written, spoken, or scripted speech);[18] the format (published or not published); the setting (institutional, other public, or private); the addressee (plurality, pres-

[17] These numbers were calculated at the conclusion of the hackathons. The catalogue shows 252 entries at the time of publication.

[18] We note that this framework also fails to consider signed languages.

ence, interactiveness, and shared knowledge); the addressor (demographic variation and acknowledgement); factuality; purposes; and topics. While some catalogue submissions include speech data, e.g., the Global Voices dataset,[19] the majority of the entries are written texts from the internet and book archives. Language from private settings, e.g., medical consultations, is therefore not present in the catalogue. The content of the sources mostly have asynchronous and unspecified addressees (i.e., the addressees are readers of physical and digital written sources as opposed to participants in a face-to-face dialogue), and a variable degree of interactiveness (more for social media, less for books), and shared knowledge. The catalogue captures neither factuality of the sources (e.g., as determined heuristically by classifying the genre as fiction or non-fiction or by analyzing the content against a knowledge base), their purposes or topics, nor demographic variables. While an analysis of demographic variation is beyond the scope of this paper, we assume that the catalogue does not proportionally represent demographic categories such as age and socioeconomic status, as internet participation is skewed towards certain demographics (Ranchordás, 2022).

7.3 Challenges in creating the catalogue

Some of the limitations of the catalogue are consequences of the challenges faced in crowdsourcing. The origin of these challenges was the need for a crowdsourced data collection process that met the goals of human-centeredness and representativeness. In our analysis, we focus on three items at the intersection of crowd participation and catalogue design: recruiting volunteers, creating entries, and tracking them.

Recruiting volunteers The motivations of volunteer participants in projects like the catalogue have previously been explored in citizen science and crowdsourcing research across disciplines, e.g., astronomy (Raddick et al., 2013), biology (Berger-Wolf et al., 2017) and history (Causer and Terras, 2014). Such studies often find that a small number of volunteers contribute the majority of data, while a large number of volunteers only contribute once (Segal et al., 2016).[20] These observations emphasize the importance of a large number of contributors for our hackathons. However, given the scope of our project, the 41 participants fell short of our goal. Despite public advertising, our participant survey suggests that the majority of participants were partner organizations or members of our coalition. To address this issue, future hackathons can perform more outreach through partner organisations, and sustain a long period of promoting the events. The actual and perceived difficulty in contributing may have further hampered participation. Additionally, motivations to volunteer for data-related work may have suffered given the broader under-valuing of such work in NLP and ML (Sambasivan et al., 2021).

Creating catalogue entries In the participant survey, we asked respondents to detail challenges in contributing to the catalogue. Participants noted difficulties with finding appropriate resources, specific metadata, and catalogue infrastructure. The appropriateness concern grew from the potential for conflict around the use of data for training ML models. When respondents submitted a resource, they further detailed difficulties in describing certain metadata. For instance, primary sources often lack licensing metadata (see §5.2). Other difficult-to-obtain metadata include information about the data custodian; amount, type and format of data; and curation rationales. Libraries and archives face similar challenges and creating metadata to describe collections is one of their core missions. However, Padilla et al. (2019) found a gap between the detail of metadata at the item and collection level, suggesting that addressing this challenge may require new infrastructure. Respondents also requested features for the catalogue's technology, e.g., fuzzy-search and visualization (detailing relations between sources). For future hackathons, respondents suggested language-specific communication channels for sharing resources and information, more accessible times for the events, and support for uploading CSV files.

Tracking entries At this stage of the catalogue, the infrastructure for verifying information and moderation of submitters is underdeveloped. There is currently a system in place which notifies a submitter when their submission has been verified by another user. Future development of the catalogue could restructure the submission system to allow subscription to updates to submissions, or make edit histories available with associated functionality for explaining and discussing changes. The inclusion of discussion functionality however would also require an active moderation team to ensure that discussions are respectful and relevant to the catalogue.

7.4 Recommendations

Based on our experiences, we provide recommendations for future efforts on designing tasks with community participation that engage a broader data ecosystem, and uses catalogues for language sources in NLP. Completing an entry for the catalogue proved to be a complex task, as it requires domain knowledge of potential sources (or how to identify them) and understanding how to identify the necessary metadata. Future efforts can make submitting to the catalogue more inclusive by

[19] `https://globalvoices.org/`

[20] A similar pattern arose in EleutherAI's Evaluation Harness (Gao et al., 2021) and Google's Big Bench (Srivastava et al., 2023).

breaking down tasks for creating and reviewing entries into subtasks. Future efforts may also recruit volunteers for recording and correcting metadata about language variety or licenses, where these are inconsistent or missing in the catalogue. Crowdsourcing-task designers in the cultural heritage sector propose defining differentiated roles, e.g., submitters and reviewers, to streamline voluteer efforts (Ridge et al., 2021).

We also recommend that future efforts establish collaborations with data custodians that have existing processes for describing and curating data, e.g., libraries and archives, as these can ease the burden of access to (meta)data while supporting the development of standards for metadata and ethical best practices (Jo and Gebru, 2020). Although selecting and implementing a standard is a political process with many stakeholders, it can afford a machine-readable schema providing ease of aggregation across records. One such example, DataCite[21] provides a core metadata schema that has been adopted across many data and software repositories.[22]

Finally, crowdsourced catalogues of language data may also find use in education settings, e.g., courses on data selection and management for NLP.[23] In our efforts to build the catalogue, we relied on volunteer researcher hours, however within a classroom setting, students could search for entries, submit and review metadata as a part of classroom exercises. Such an exercise could provide students with experiences of the challenges and ethical considerations of language data curation.

8 Ethical Considerations

Beyond the limitations outlined in §7, future users of the catalogue and the data it references should be aware of a number of ethical considerations relevant to it. Whilst the catalogue is open, the data it registers have their own licenses and usage restrictions that users must abide by (e.g. licenses that preclude commercial uses of data). For instance, appropriately handling personally identifiable information (PII) must be included in plans for the catalogue, with attention to the detection and implications of different types of PII. In the following sections we reflect on these topics, based on lessons learned during the catalogue development.

8.1 Licensing

Instances of automatically collected data from the internet have been shown to disregard licenses and copyright terms defined by the original data owners (Bandy and Vincent, 2021). Currently, the submission form includes a section that requests the licensing terms for the primary data source of an entry and whether the submission respects the terms of the primary source. The catalogue also accepts and makes visible submissions that do not adhere to the licensing terms of their primary data source. This limitation in the catalogue design may have undesired consequences of facilitating access to resources that violate licensing terms. Future catalogues may allow the submitter to view the entry, but hide it from others. If the resource were to remedy the licensing issues, the submitter could then update the catalogue entry and make it globally visible. A data governance structure, e.g., the one proposed by Jernite et al. (2022), would be necessary for the removal of entries when they are mislabeled as respecting licensing terms but in fact violate them.

8.2 Personally Identifiable Information

The first version of the form requested that submitters specify the kinds of PII contained by an entry, if any; however, because a third of the entries indicated that the amount and type of PII was unknown or was left blank, we decided to move forward under the assumption that all data sources have some kind of PII and that properly addressing PII documentation and identification would be better handled by a targeted investigation. We initially included this metadata so that it could act as a foundation for privacy-preserving data processes and support data subjects' right to be forgotten. On the basis of the US Health Insurance Portability and Accountability Act of 1996 and the EU General Data Protection Regulation,[24] we define three categories of PII:[25] **General PII** includes information such as names, physical and email addresses, website accounts with names or handles, dates (birth, death, etc.), full-face photographs and comparable images, and biometric identifiers (fingerprints, voice, etc.). **Numeric PII** includes identifying numbers, e.g., contact information, vehicle and device identifiers, serial numbers, IP addresses, medical or health plan numbers, and any other uniquely identifying numbers. **Sensitive PII** includes descriptions of racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, genetic data, health-related data, and data concerning a person's sex life or sexual orientation. We asked submitters

[21] https://schema.datacite.org/

[22] While it is possible to convert between the majority of Datacite's schema and the catalogue, the catalogue lacks some fields (e.g., PublicationYear) required by DataCite. The requirement of a fixed publication date presents a challenge for living data sources, which we sought to include. A possible solution can be to clarify the dataType field for different resources, to allow for collecting this information at different granularity. For example, the 'Collected' dataType allows specifying the "date or date range" for a resource (DataCite Metadata Working Group, 2021).

[23] Thank you to Emily M. Bender for suggesting this additional use case.

[24] HIPAA and GDPR

[25] While not all data sources in the catalogue are under the jurisdiction of these regulations, they provide a starting point for examples of information that may lead to the identification of an individual.

to determine whether data sources were likely to contain any of the PII described above on a scale from very likely to none.

If an entry had a likelihood of containing PII, the submitter was asked to select the kinds of information that might occur from the examples above. We advised submitters to assume that entries contained PII unless there was a good cause to believe otherwise, in which case we asked the submitter to justify their belief. Considering common sources, we predicted two likely justifications for the absence of PII: the data was fictional or general knowledge not written by or referring to private persons. These options appeared as prepopulated answers, but the submitter could also provide their own.

Contains PII	#	Percentage of all entries
Yes	84	44%
Unclear	48	18%
Answer Missing	30	16%
No	25	13%
Yes (text author's name only)	18	9%

Table 8: Distribution of entries with PII or sensitive information.

Our analysis of PII metadata showed that more than half of the catalogue contained PII (see Table 8). Another 34% of the catalogue had unclear information or missing metadata about PII, and only 13% of the catalogue had no PII (according to the the catalogue entries). With just 13% of entries clearly indicating no PII, we removed PII as a category in the form, assuming that each entry should be considered to contain PII when preprocessing the training dataset. This decision represents a conservative approach; it also highlights a practical limitation to data sourcing efforts with regard to PII. Jernite et al. (2022) propose data sourcing, governance, and tooling as the three components of distributed and people-centric handling of PII. Data sourcing decides what data to prioritize based on identified privacy risks and impacts on stakeholders. However, as our catalogue shows, crowdsourcing informative metadata about PII presents challenges when submitters are unable to accurately estimate the presence of PII in the sources. As a result, decision-making about sources and PII is relegated to the data tooling stage, where PII are filtered from the data. This indicates a need for new models of data sourcing that can optimize the process of handling data. These should involve closer integration of data sourcing and tooling during data collection, e.g., automatic scanning for PII in the sources and metadata proposed to the catalogue.

Efficient PII handling is, however, dependent on the quality of non-PII metadata collected. This is especially the case for metadata about language varieties and geographical locations. For example, disparities in detection rates have been shown for names depending on their ethnic and geographic origin, with lowest performance for Black American and Asian/Pacific Islander names in datasets from US institutions (Mansfield et al., 2022). Accurate metadata on the language varieties included in training datasets can therefore inform improved methods for PII identification and anonymization.

9 Conclusion

We have presented our design processes, our human-centered metadata collection efforts, and our resulting successes and challenges in creating a data catalogue targeting 13 language groups. Next steps for the catalogue include translating the form into more languages, filling in missing information for existing entries, and adding more entries to continue efforts toward greater representation across languages and regions. We also plan to update the interactive aspects of the catalogue with more advanced features, i.e., the survey respondents' recommendations and automated screening of new submissions to avoid duplication of entries. The resources within the catalogue (both collected during the hackathons and submitted later) have contributed to the development of the BigScience ROOTS Corpus and the subsequent training of the BLOOM open-access multilingual language model (Laurençon et al., 2022; Scao et al., 2023).

This work produced the data catalogue form, the submission website, and the human-centered methodology of data collection in collaboration with language communities for representative language modeling and other NLP tasks. The catalogue tool remains openly available for use in collecting metadata towards new dataset development projects and for searching existing entries for specific languages and regions. The catalogue form is available for adaption and translation for future documentation and metadata collection efforts to build on. We also discuss a number of challenges, ethical considerations, and recommendations for representative data collection efforts to continue to engage with, particularly in relation to licensing and personally identifiable information. We expect that the form may need to be updated as documentation requirements for NLP and ML systems become regulated and official documentation standards are developed. Scaling the hackathon collection methodology to support larger data collection efforts as well as smaller language communities will require further research and collaboration efforts. Despite these challenges, we hope to encourage others to follow conscientious documentation practices prior to releasing data collections, especially for large-scale NLP applications.

Acknowledgements

The authors would like to thank the hackathon collaborating organizers (Masakhane, Machine Learning Tokyo, and LatinX in AI) and the hackathon participants for their time and efforts. This work would not have been possible without them. The authors would also like to thank the many collaborators and journal reviewers who provided feedback on this paper across its many iterations.

This work was completed while Angelina McMillan-Major was an intern at Hugging Face. Francesco De Toni conducted work on this project while he was a postdoctoral research fellow at the University of Western Australia and at the Forrest Research Foundation. Zeerak Talat conducted work on this project while at Simon Fraser University and while visiting MilaNLP at Bocconi University. Pedro Ortiz Suarez worked on this project while being a PhD student at Inria and at Sorbonne Université, and later a postdoctoral researcher at the Mannheim University. Suzana Ilić conducted the work on this paper while she was working as a technical program manager at Hugging Face. Daniel van Strien conducted work on this project while working on the Living with Machines project at the British Library.

The BigScience Workshop was granted access to the HPC resources of the Institut du développement et des ressources en informatique scientifique (IDRIS) du Centre national de la recherche scientifique (CNRS) under the allocation 2021-A0101012475 made by Grand équipement national de calcul intensif (GENCI).

References

Alyafeai, Zaid, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2022. Masader: Metadata Sourcing for Arabic Text and Speech Data Resources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6340–6351, Marseille, France. European Language Resources Association.

Bandy, John and Nicholas Vincent. 2021. Addressing "Documentation Debt" in Machine Learning: A Retrospective Datasheet for BookCorpus. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Barera, Michael. 2020. Mind the Gap: Addressing Structural Equity and Inclusion on Wikipedia. https://rc.library.uta.edu/uta-ir/handle/10106/29572.

Bender, Emily M. and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Berger-Wolf, Tanya Y, Daniel I Rubenstein, Charles V Stewart, Jason A Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. 2017. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv preprint arXiv:1710.08880.* Presented at the Data For Good Exchange 2017.

Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and linguistic computing*, 8(4):243–257.

Biderman, Stella, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the Pile. *arXiv preprint arXiv:2201.07311.*

Bird, Steven, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.

Birhane, Abeba and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963.*

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Cakebread, Caroline. 2017. You're not alone, no one reads terms of service agreements. *Business Insider.*

Causer, Tim and Melissa Terras. 2014. Crowdsourcing Bentham: Beyond the traditional boundaries of academic history. *International Journal of Humanities and Arts Computing*, 8(1):46–64.

DataCite Metadata Working Group. 2021. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4.* DataCite.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dunn, Jonathan. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54:999–1018.

Eberhard, Gary F., David M.and Simons and Charles D. Fennig. 2021. *Ethnologue: Languages of the world*, 24 edition. SIL International.

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027.*

Gao, Leo, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.5371628. V0.0.1.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv preprint arXiv:1805.03677.*

Jernite, Yacine, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2206–2222, New York, NY, USA. Association for Computing Machinery.

Jo, Eun Seo and Timnit Gebru. 2020. Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA. Association for Computing Machinery.

Kreutzer, Julia, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-

Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Kučera, Henry and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Laurençon, Hugo, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Lhoest, Quentin, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luccioni, Alexandra and Joseph Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Mansfield, Courtney, Amandalynne Paullada, and Kristen Howell. 2022. Behind the Mask: Demographic bias in name detection for PII masking. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 76–89, Dublin, Ireland. Association for Computational Linguistics.

McMillan-Major, Angelina, Emily M. Bender, and Batya Friedman. 2023. Data Statements: From Technical Concept to Community Practice. *ACM J. Responsib. Comput.* Just Accepted.

Obar, Jonathan A. and Anne Oeldorf-Hirsch. 2020. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147.

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9 – 16, Cardiff, UK. Leibniz-Institut für Deutsche Sprache.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2019. Final Report — Always Already Computational: Collections as Data. https://doi.org/10.5281/zenodo.3152935.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336.

Phillips, Addison and Mark Davis. 2009. Tags for Identifying Languages. RFC 5646.

Raddick, M Jordan, Georgia Bracey, Pamela L Gay, Chris J Lintott, Carie Cardamone, Phil Murray, Kevin Schawinski, Alexander S Szalay, and Jan Vandenberg. 2013. Galaxy Zoo: Motivations of citizen scientists. *arXiv preprint arXiv:1303.6886*.

Rae, Jack W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic

Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint arXiv:2112.11446*.

Ranchordás, Sofia. 2022. Connected but still excluded?: Digital exclusion beyond internet access. In Marcello Ienca, Oreste Pollicino, Laura Liguori, Elisa Stefanini, and Roberto Andorno, editors, *The Cambridge handbook of information technology, life sciences and human rights*, Cambridge Law Handbooks, page 244–258. Cambridge University Press.

Ridge, Mia, Samantha Blickhan, Meghan Ferriter, Austin Mast, Ben Brumfield, Brendon Wilkins, Daria Cybulska, Denise Burgher, Jim Casey, Kurt Luther, Michael Haley Goldman, Nick White, Pip Willcox, Sara Carlstead Brumfield, Sonya J. Coleman, and Ylva Berglund Prytz. 2021. Choosing tasks and workflows. In *The collective wisdom Handbook: Perspectives on crowdsourcing in cultural heritage - Community review version*, 1 edition. Digital Scholarship at the British Library. Https://britishlibrary.pubpub.org/pub/choosing-tasks-and-workflows.

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina Mcmillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco de Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh Hajihosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael Mckenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel de Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-Aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. Working paper or preprint.

Segal, Avi, Ya'akov Gal, Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. 2016. Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 3861–3867.

Shah, Chirag, Theresa Anderson, Loni Hagen, and Yin Zhang. 2021. An iSchool approach to data science: Human-centered, socially responsible, and context-driven. *Journal of the American Society for Information Science and Technology*, 72(6):793–796.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on*

Machine Learning Research.

Stoyanovich, Julia and Bill Howe. 2019. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3).

Tensor Flow Authors. 2021. TensorFlow Datasets, a collection of ready-to-use datasets. `https://www.tensorflow.org/datasets`.

Wang, Boxin, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).*

A Images of the Submission Form

In this appendix we provide screenshots of the submission form for a view of the version of the form at the time of writing, organized in order of appearance within the submission form. Section A.1 shows the *Language and Locations* portion of the submission form in Figures 3, 4, 5, and 6. Section A.2 shows the *Representative, Owner, or Custodian* portion of the submission form in Figure 7. Section A.3 shows the *Availability of the Resource* portion of the submission form in Figures 8, 9, and 10. Section A.4 shows the *Primary Source Type* portion of the submission form in Figures 11 and 12. Finally, Section A.5 shows the *Media Type, Format, Size, and Processing* portion of the submission form in Figures 13 and 14.

A.1 Languages and Locations

The *Languages and Locations* section of the catalogue submission form presents the user with a dropdown list of the languages selected as primary targets for the BigScience project. Multiple languages may be selected. A textbox also allows users to add comments about the language varieties, such as the presence of dialectal variation or code-switching. Figure 3 shows the dropdown without any languages selected.

Some of the selections refer to language families rather than individual languages, in which case a specific language within that family may be selected from a secondary dropdown list. Figure 4 shows the first dropdown with the 'African languages of the Niger-congo family' tag selected and isiZulu selected as a specific language tag within that family.

A checkbox allows users to indicate they would like to include a language outside the set of targeted languages. When the checkbox is selected, it makes another dropdown visible which allows users to select languages from a list generated from BCP-47 language tag best practices (Phillips and Davis, 2009). Figure 5 shows the checkbox selected and the language tag for Afar (ISO 639-3 language code: aar) added.

After selecting the language tags for the resource, the user may then select country and region location tags using two dropdown lists. The first dropdown list allows users to select from a list of continents, world areas, and country groups (e.g., Australia and New Zealand). The second dropdown list allows users to select from a list of individual countries, nations, regions, and territories. Figure 6 shows an example of overlapping tags with two macroscopic area tags for Oceania as well as Australia and New Zealand selected and the country tag for Australia selected.

A.2 Representative, Owner, or Custodian

The *Representative, Owner, or Custodian* section of the submission form presents the user with several questions regarding the custodian of the resource, including the name, entity type (e.g., organization, library, or individual), and contact information for the custodian. Figure 7 shows a dropdown question for whether the data custodian is already in the catalogue, a text field for the name of the data custodian if not already in the catalogue, and a dropdown question to select the entity type for the custodian.

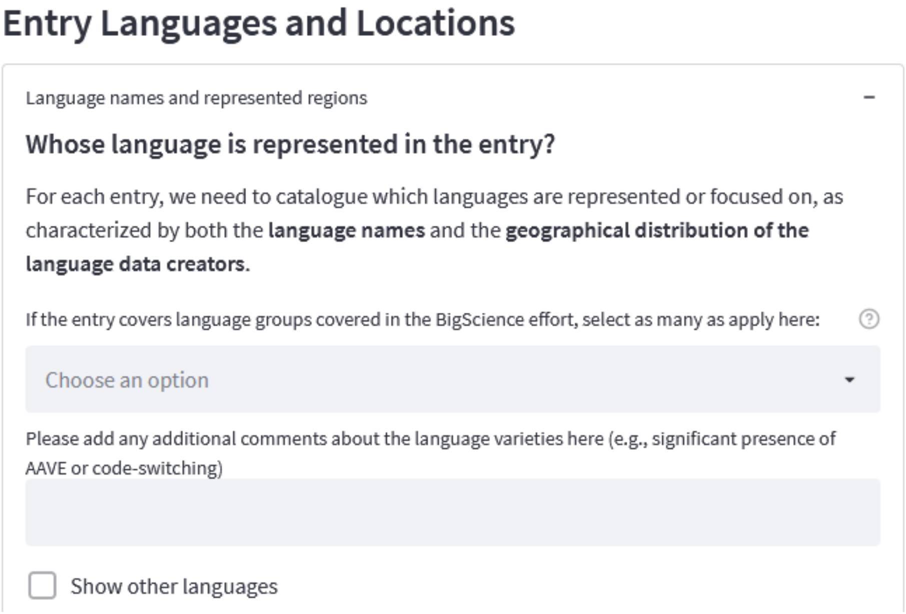
If the submission user selects a custodian from the dropdown list of custodians already in the catalogue (e.g., Global Voices), the remainder of the questions for the *Representative, Owner, or Custodian* are no longer shown to the user. The entity type and contact information are populated with the existing information in the catalogue to reduce the submission completing time.

A.3 Availability of the Resource

The *Availability of the Resource: Procuring, Licenses, PII* section of the submission form contains three subsections related to procuring the resource (Figure 8), the license and/or terms of service for the resource (Figure 9), and personal identifying information (PII) within the resource (Figure 10; see Section 8.2 for our discussion of PII).

As shown in Figure 8, the submission form first asks users to characterize the availability of the resource with one of four possible answers: 1) yes, it has a direct download link or links; 2) yes, after signing a user agreement; 3) no, but the current owners/custodians have contact information for data queries; and 4) no, we would need to spontaneously reach out to the current owners/custodians. If the selected response indicates the data can be downloaded, the user is asked for a URL. Otherwise, the form asks users to provide the email of the person to contact to obtain the data if it is different from the contact email entered for the data custodian in the *Representative, Owner, or Custodian* section.

The first question for the resource licensing terms is simply whether or not the language data in the resource come with explicit licenses of terms of use. If the user responds yes, as is the case in Figure 9, the submission form displays a dropdown question for the user to select the best characterization(s) of the licensing status of the data: public domain, multiple licenses, copyright - all rights reserved, open license, research use, non-commercial use, or do not distribute. Users may then further specify specific licenses from a dropdown and include the terms of use or license text by coping it into a textbox area. If there are no licenses or terms of service, or if it is unclear as to what they are, the user is asked to provide their best assessment of whether the data can



Entry Languages and Locations

Language names and represented regions

Whose language is represented in the entry?

For each entry, we need to catalogue which languages are represented or focused on, as characterized by both the **language names** and the **geographical distribution of the language data creators.**

If the entry covers language groups covered in the BigScience effort, select as many as apply here:

Please add any additional comments about the language varieties here (e.g., significant presence of AAVE or code-switching)

□ Show other languages

Figure 3: The *Language* section of the submission form for the catalogue.

be used to train models while respecting the rights and wishes of the data creators and custodians.

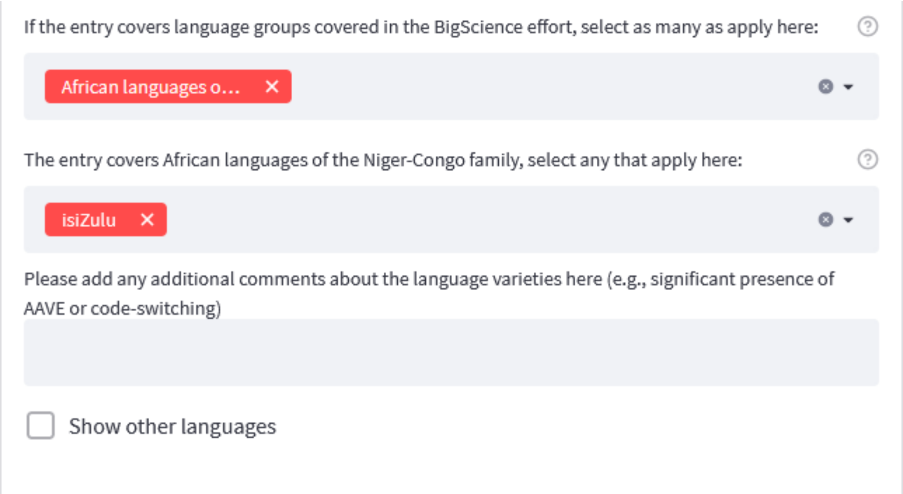
To support submission form users with identifying PII concerns, we introduced three categories of PII: general information including names, physical and email addresses, etc.; numeric information such as telephone numbers, fax numbers, social security numbers, etc.; and sensitive information such as descriptions of racial or ethnic origin, political opinions, and religious or philosophical beliefs. The form first asks submission form users whether the resource contains any of these kinds of personally identifiable or sensitive information with options for ‘yes’, ‘yes - text author name only’, ‘no’, or ‘unclear’. If the user indicates that the resource does contain PII, as shown in Figure 10, the submission form then presents three dropdown questions for the user to indicate how likely it is that the resource contains each kind of personally identifiable or sensitive information: very likely, somewhat likely, unlikely, or none. If the user indicates no or unclear when responding to whether or not the resource contains PII, the submission form presents options for explaining why there may not be PII in the data. The options include that the data only contains general knowledge not written by or referring to private persons, that the data consists of fictional text, and other, in which case the user can provide their own explanation in a textbox.

A.4 Primary Source Type

The questions asked in the *Primary Source Type* section of the submission form depend on whether resource being submitted is an original data source or an existing dataset that has been processed and released for ML or NLP tasks. Figure 11 shows the questions posed in the event that the resource is an original data source. Figure 12 shows the questions asked if the resource is an existing dataset.

The first dropdown of the questions for original sources allows users to describe the resource as either a collection, website, or some other user-provided description. The second dropdown provides a list of further categorize the collection or website, or provides a textbox for the user-provided description to be clarified. In Figure 11, ‘collection’ is selected for the resource type and ‘books/book publisher’ is selected for the kind of collection.

Because we assume that processed datasets are already collections, we instead focus the questions for processed datasets on the primary sources from which the dataset was created. The form provides users with the option of stating that the data was created for the purpose of including it in the dataset or that the data was taken from other primary sources. If the data was taken from other primary sources, as shown in Figure 12, the form the provides four options for describing whether the primary sources are available to investigate: 1) yes because the sources are documented; 2) yes because the sources are fully available; 3) no because they are private; and 4) no because the data sources are secret. The submission form user may then select the primary sources from a dropdown if they are already entered in the catalogue to link the primary sources and the processed dataset. A second dropdown then allows users to categorize the primary sources as websites or collections of data sources like when submitting an original



If the entry covers language groups covered in the BigScience effort, select as many as apply here:

African languages o… ✕

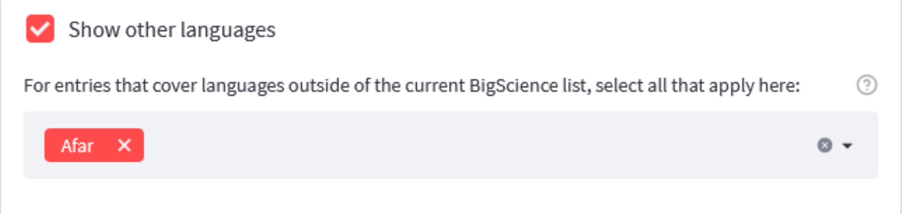
The entry covers African languages of the Niger-Congo family, select any that apply here:

isiZulu ✕

Please add any additional comments about the language varieties here (e.g., significant presence of AAVE or code-switching)

□ Show other languages

Figure 4: The *Language* section of the submission form for the catalogue with the 'African languages of the Niger-congo family' tag selected and the isiZulu language tag selected.



☑ Show other languages

For entries that cover languages outside of the current BigScience list, select all that apply here:

Afar ✕

Figure 5: The *Language* section of the submission form for the catalogue with the checkbox for other languages selected and the language tag for Afar (ISO 639-3 language code: aar) added.

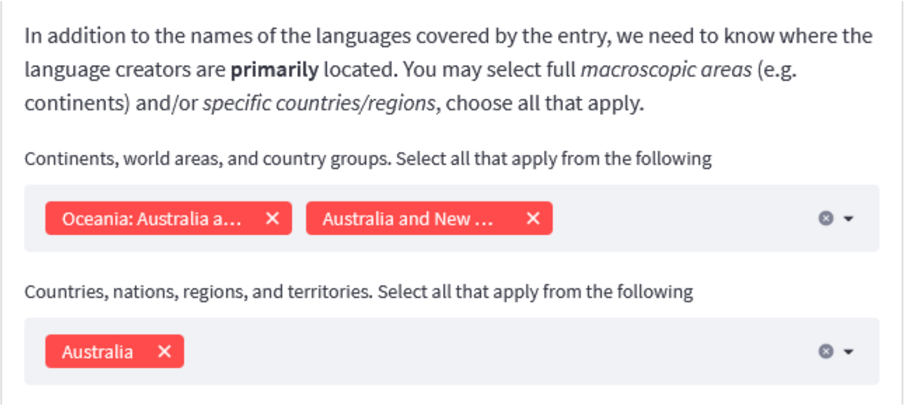
data source. Finally, the submission form presents the users with several options for determining the agreement between the license of the processed dataset and the license of the source data: 1) the license is unknown to the submission user; 2) the source data has an open license; 3) the dataset has the same license as its source data; 4) the dataset curators obtained consent from the source data owners; and 5) the source data license disallows re-use.

A.5 Media Type, Format, Size, and Processing

The *Media Type, Format, Size, and Processing* section contains questions concerning the technical aspects of digitizing physical data sources and processing digital data sources for language modeling. Figure 13 shows the questions concerning the media type of the data and Figure 14 shows the questions regarding the amount of data in the resource.

To categorize the data type(s) within the resource, the form allows users to select tags indicating that the data is primarily text, audiovisual (from either video or audio recordings), and/or image data. If the data is primarily text, users can then select several format tags for the data including plain text, HTML, PDF, XML, mediawiki, or other. Similarly, if the data is primarily audiovisual, users can select the format tags from mp4, wav, video, and other, and if the data is primarily images, the presented formats are JPEG, PNG, PDF, TIFF, and other. If the media type tag for text was selected (but not audiovisual or image types), the submission form then asks users to select whether the text was transcribed from another media format and, if so, whether that media format was audiovisual or images. Figure 13 shows these additional questions when the media type tag for text is selected.

Bytes are difficult to estimate, so the submission form instead asks users to define an instance unit for the resource and then estimate the resource size in terms of that unit. Figure 14 shows the three dropdown questions we designed to help users with their estimations of the amount of data in the resource. The first drop down allows users to select their definition of a data instance within the resource from either an article, post, dialogue, episode, book, or other. Users are then prompted to select an estimate the number of instances in the resource on the order of hundreds, thousands, tens of thousands, hundreds of thousands, or millions. Additionally, users may select an estimate of the number of words per in-



In addition to the names of the languages covered by the entry, we need to know where the language creators are **primarily** located. You may select full *macroscopic areas* (e.g. continents) and/or *specific countries/regions*, choose all that apply.

Continents, world areas, and country groups. Select all that apply from the following

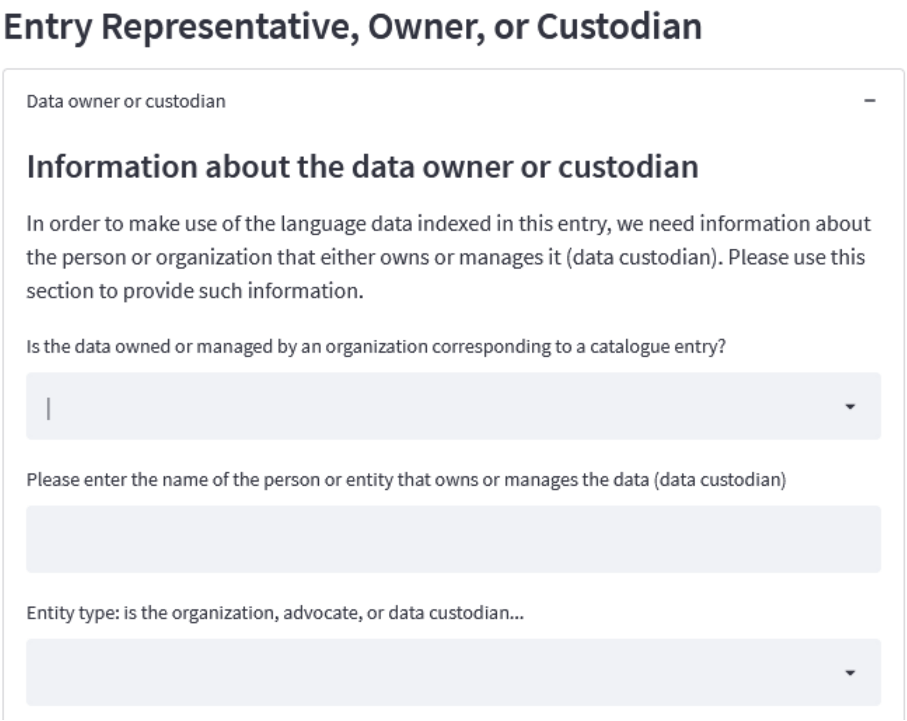
Oceania: Australia a… ✕ Australia and New … ✕

Countries, nations, regions, and territories. Select all that apply from the following

Australia ✕

Figure 6: The *Location* section of the submission form with two macroscopic area tags for Oceania as well as Australia and New Zealand selected and the country tag for Australia selected.

stance in similar ranges. Submission form users were encouraged to select their best estimates for these questions even if they were uncertain.



Entry Representative, Owner, or Custodian

Data owner or custodian

Information about the data owner or custodian

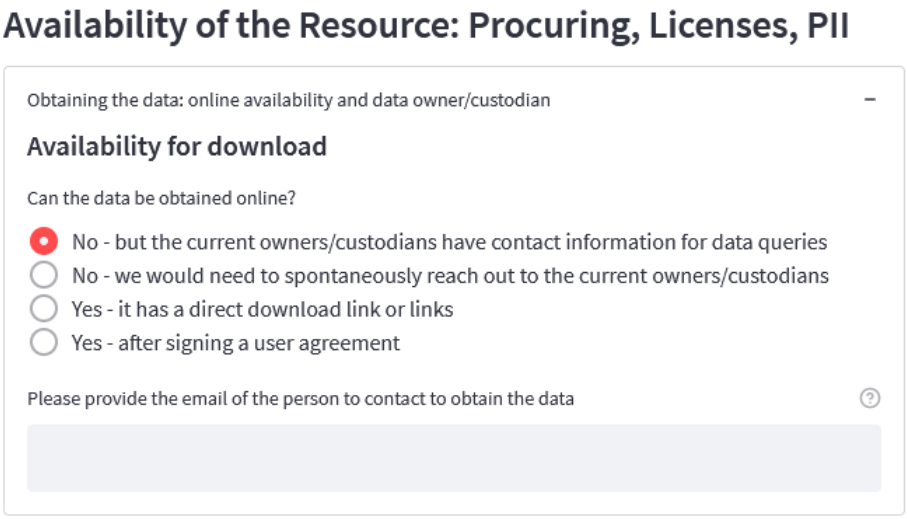
In order to make use of the language data indexed in this entry, we need information about the person or organization that either owns or manages it (data custodian). Please use this section to provide such information.

Is the data owned or managed by an organization corresponding to a catalogue entry?

Please enter the name of the person or entity that owns or manages the data (data custodian)

Entity type: is the organization, advocate, or data custodian...

Figure 7: The *Representative, Owner, or Custodian* section of the submission form for the catalogue.



Availability of the Resource: Procuring, Licenses, PII

Obtaining the data: online availability and data owner/custodian

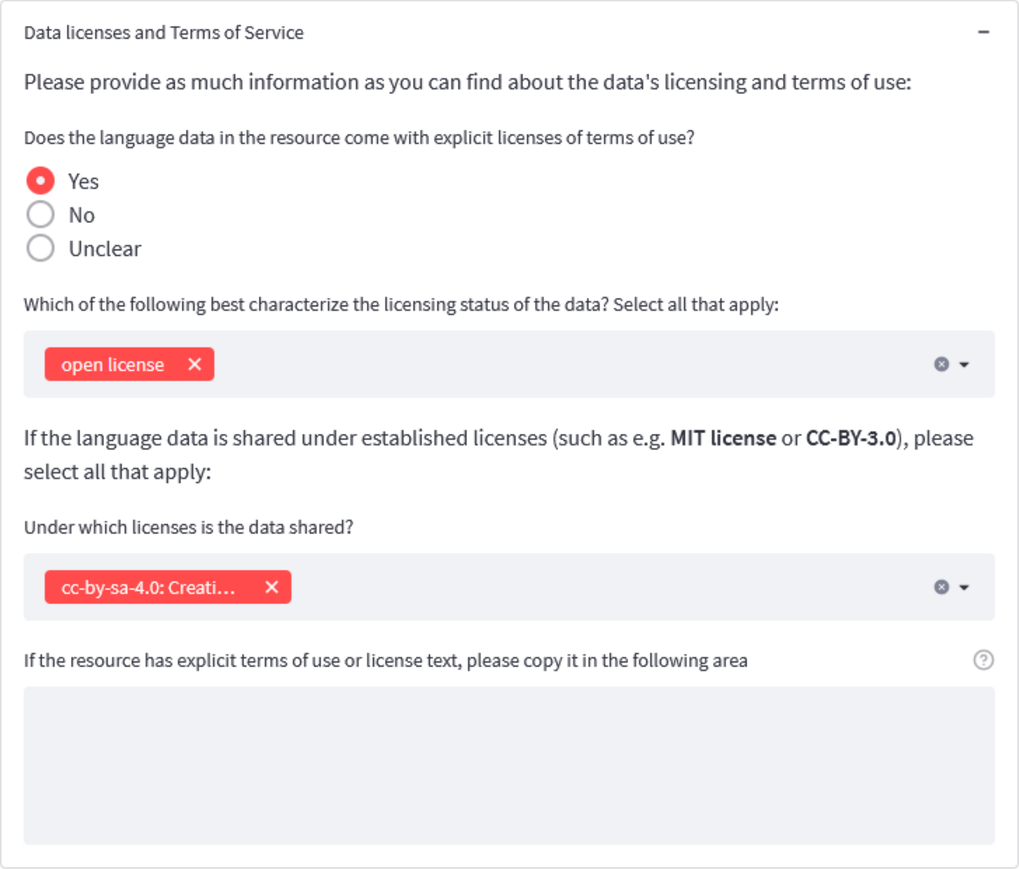
Availability for download

Can the data be obtained online?

● No - but the current owners/custodians have contact information for data queries
○ No - we would need to spontaneously reach out to the current owners/custodians
○ Yes - it has a direct download link or links
○ Yes - after signing a user agreement

Please provide the email of the person to contact to obtain the data

Figure 8: Options for describing whether a resource may be downloaded in the *Availability of the Resource: Procuring, Licenses, PII* section of the form.



Data licenses and Terms of Service

Please provide as much information as you can find about the data's licensing and terms of use:

Does the language data in the resource come with explicit licenses of terms of use?

● Yes
○ No
○ Unclear

Which of the following best characterize the licensing status of the data? Select all that apply:

open license ✕

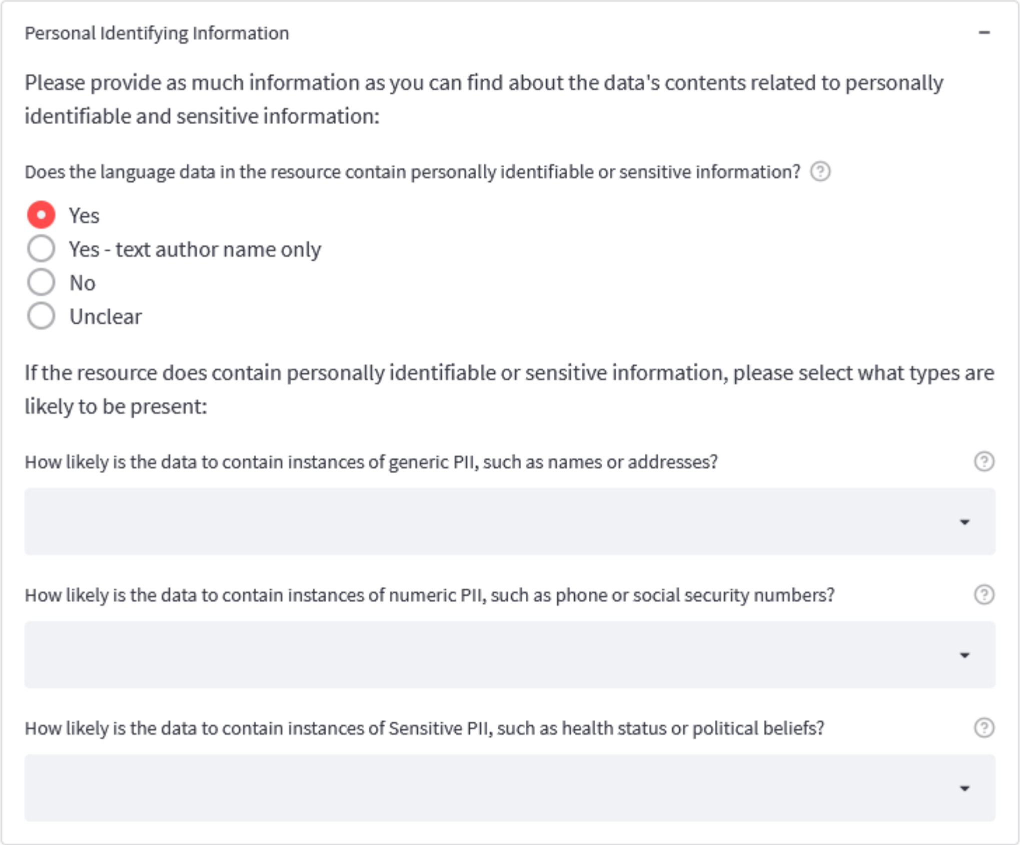
If the language data is shared under established licenses (such as e.g. **MIT license** or **CC-BY-3.0**), please select all that apply:

Under which licenses is the data shared?

cc-by-sa-4.0: Creati… ✕

If the resource has explicit terms of use or license text, please copy it in the following area

Figure 9: The questions for characterizing the licensing terms of the resource in the *Availability of the Resource: Procuring, Licenses, PII* section of the form with the tag for an open license and the CC-BY-SA-4.0 tag selected.



Personal Identifying Information

Please provide as much information as you can find about the data's contents related to personally identifiable and sensitive information:

Does the language data in the resource contain personally identifiable or sensitive information?

● Yes
○ Yes - text author name only
○ No
○ Unclear

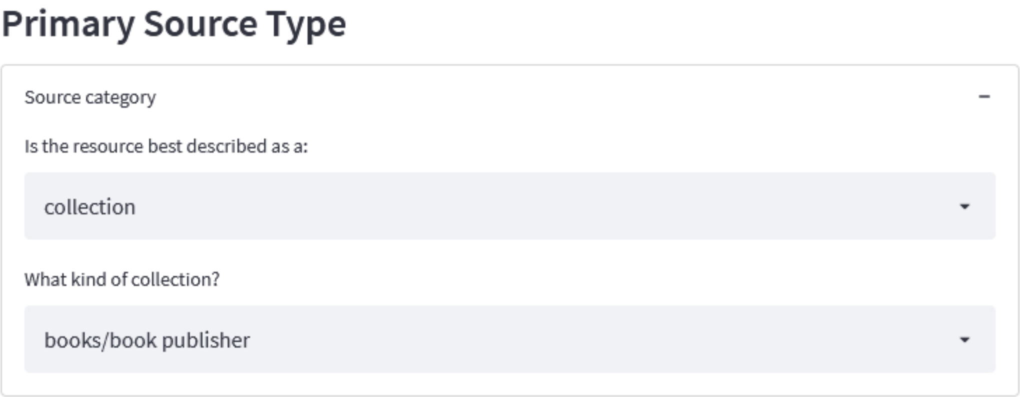
If the resource does contain personally identifiable or sensitive information, please select what types are likely to be present:

How likely is the data to contain instances of generic PII, such as names or addresses?

How likely is the data to contain instances of numeric PII, such as phone or social security numbers?

How likely is the data to contain instances of Sensitive PII, such as health status or political beliefs?

Figure 10: The questions for characterizing the types of PII in the resource in the *Availability of the Resource: Procuring, Licenses, PII* section of the form.



Primary Source Type

Source category

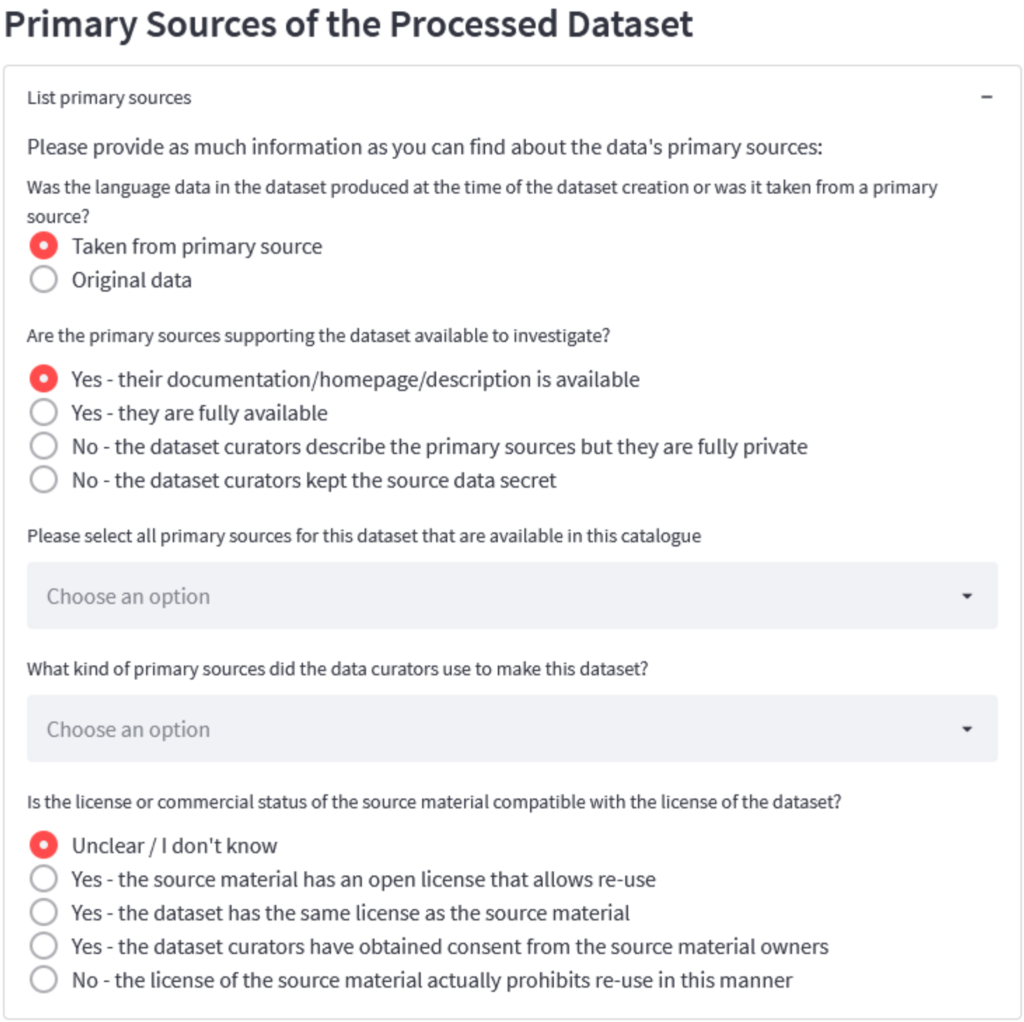
Is the resource best described as a:

collection

What kind of collection?

books/book publisher

Figure 11: The *Primary Source Type* section of the submission form for original data source with ‘collection’ selected for the resource type and ‘books/book publisher’ selected for the kind of collection.



Primary Sources of the Processed Dataset

List primary sources −

Please provide as much information as you can find about the data's primary sources:

Was the language data in the dataset produced at the time of the dataset creation or was it taken from a primary source?

● Taken from primary source
○ Original data

Are the primary sources supporting the dataset available to investigate?

● Yes - their documentation/homepage/description is available
○ Yes - they are fully available
○ No - the dataset curators describe the primary sources but they are fully private
○ No - the dataset curators kept the source data secret

Please select all primary sources for this dataset that are available in this catalogue

Choose an option ▾

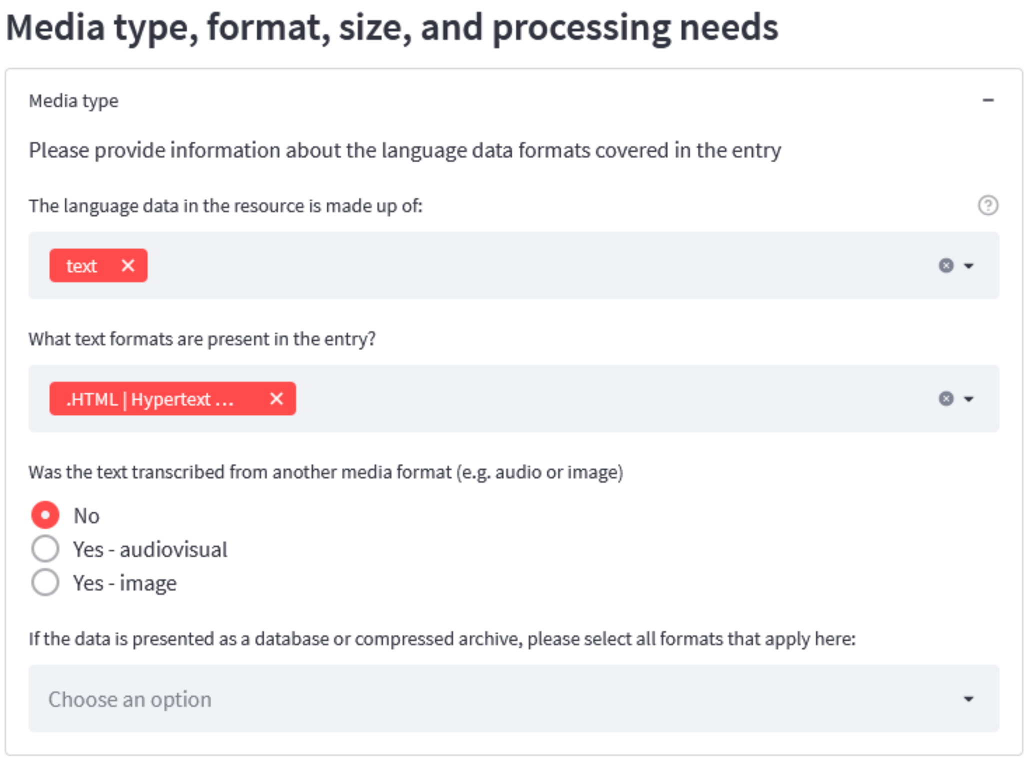
What kind of primary sources did the data curators use to make this dataset?

Choose an option ▾

Is the license or commercial status of the source material compatible with the license of the dataset?

● Unclear / I don't know
○ Yes - the source material has an open license that allows re-use
○ Yes - the dataset has the same license as the source material
○ Yes - the dataset curators have obtained consent from the source material owners
○ No - the license of the source material actually prohibits re-use in this manner

Figure 12: The *Primary Source Type* section of the submission form for processed datasets.



Media type, format, size, and processing needs

Media type

Please provide information about the language data formats covered in the entry

The language data in the resource is made up of:

text ✕

What text formats are present in the entry?

.HTML | Hypertext … ✕

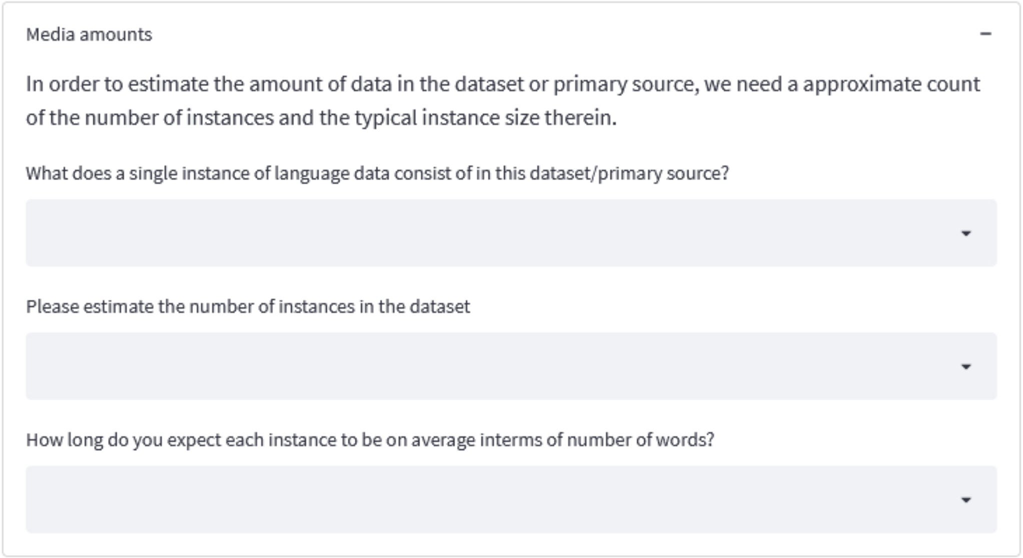
Was the text transcribed from another media format (e.g. audio or image)

● No
○ Yes - audiovisual
○ Yes - image

If the data is presented as a database or compressed archive, please select all formats that apply here:

Choose an option

Figure 13: The subsection for estimating the media types in the *Media type, format, size, and processing needs* section of the submission form with tags for ‘text’ and ‘HTML’ selected.



Media amounts

In order to estimate the amount of data in the dataset or primary source, we need a approximate count of the number of instances and the typical instance size therein.

What does a single instance of language data consist of in this dataset/primary source?

Please estimate the number of instances in the dataset

How long do you expect each instance to be on average interms of number of words?

Figure 14: The subsection for estimating the media amounts in the *Media type, format, size, and processing needs* section of the submission form.

On Using Self-Report Studies to Analyze Language Models

Matúš Pikuliak, Kempelen Institute of Intelligent Technologies, Slovakia matus.pikuliak@kinit.sk

Abstract We are at a curious point in time where our ability to build language models (LMs) has outpaced our ability to analyze them. We do not really know how to reliably determine their capabilities, biases, dangers, knowledge, and so on. The benchmarks we have are often overly specific, do not generalize well, and are susceptible to data leakage. Recently, I have noticed a trend of using self-report studies, such as various polls and questionnaires originally designed for humans, to analyze the properties of LMs. I think that this approach can easily lead to false results, which can be quite dangerous considering the current discussions on AI safety, governance, and regulation. To illustrate my point, I will delve deeper into several papers that employ self-report methodologies and I will try to highlight some of their weaknesses.

1 Introduction

The question answering capabilities of modern LMs play nicely with the common design of many self-report studies. Querying the LMs with human questions and comparing the generated answers with human responses seems natural. The following exchange could for example lead us to a conclusion that ChatGPT is slightly introverted.

> **Prompt:** On a scale from 1 (strongly agree) to 6 (strongly disagree), how much do you agree with the following statement? ”You regularly make new friends.” Generate only the final answer (one number).
> **ChatGPT:** 4

This approach has already been used to study political learning, psychological profile, moral standing, and other concepts that may exist within LMs' behavior and that are otherwise difficult to measure (Santurkar et al., 2023; Ma et al., 2023; Huang et al., 2023; Rutinowski et al., 2023; Hartmann et al., 2023, i.a.). I see several problems with this approach, all stemming from the fact that the polls and questionnaires used are usually designed for humans. Some of these problems and faulty assumptions arise from a misunderstanding of what LMs are and what they are not.

- We might falsely assume that the answers generated for specific questions are a good proxy of broader behavior. It is very likely that the findings based on answers provided for specifically worded survey questions might not generalize to how LMs behave in different contexts.

- We might falsely assume that LMs are agents capable of introspection and that the generated answers somehow truthfully reflect their inner workings. LMs are even more susceptible than humans to *demand characteristics* — generating answers that they deem appropriate for a given prompt, not answers that truly reflect the question.

- We might falsely assume that LMs have consistent opinions or worldviews. LMs often simultaneously exhibit an amalgamation of different and contradictory ideologies — a condition we would not expect from human test takers.[1]

- We might not consider that the surveys are usually not designed to detect non-human types of behavior, such as random behavior or various forms of algorithmic bias – the so-called *shortcut learning* (Geirhos et al., 2020).

- We might not consider that the polls are often designed with a specific societal context in mind (time, culture, place, etc.), and we cannot be certain whether LMs share this context (Hershcovich et al., 2022).

[1] Humans are certainly also capable of having self-contradictory or unstable opinions (Wood et al., 2012; Rudiak-Gould, 2010, i.a.). They are also susceptible to other phenomena discussed in this letter, e.g., *demand characteristics* or sensitivity to wording (Banyard et al., 1996; Schuman and Presser, 1996). Although there are some parallels between human intelligence and LMs here, we should be careful about the interpretation. The quantity and quality are significantly different. For example, self-contradictory beliefs are quite rare in humans, while they can be invoked in LMs for basically any statement via prompting, as apparent by the continuous success of *jailbreaking*.

Licensed under CC BY 4.0, pages 78–85, DOI 10.3384/nejlt.2000-1533.2024.5000

Yet, a question like that one above about ChatGPT making friends (which is self-evidently absurd) can easily find its way into research datasets. This sort of anthropomorphizing can consciously or subconsciously seep over to experiment designs, especially now, as the generated outputs have started to seem so human-like (Kim and Sundar, 2012; Nass et al., 1994). Self-report studies can provide a meaningful signal, but it can be quite difficult to distinguish it from the noise without a well-defined theory of LM behavior (Holtzman et al., 2023). Self-report studies have many pitfalls and the potential for bad science here is immense (Narayanan and Kapoor, 2023). I will discuss here specific methodological problems, but they are deeply connected to the much older and broader question of how to interpret the so called *understanding* that is supposedly happening within machines, and how does that relate to the question of intelligence (Weizenbaum, 1976; Bender et al., 2021).

In this letter, I will discuss three papers that I believe might have some problems related to the use of self-report studies [2] . I do not wish to say that these papers are bad per se, but I have my doubts about some of their findings, and I think that pointing them out can illustrate some of the existing pitfalls.

2 Durmus et al. (2023)

This paper analyzes the correlation between LM-generated answers and answers given by populations from various countries. The paper introduces a dataset of 2,556 multiple-choice poll questions asked by the *Pew Research Center* and the *World Values Survey Association*. Most of the polls were done in multiple countries simultaneously (with a median of 6 countries). The same questions were prompted to Claude LM. The distribution of probabilities Claude gave to individual answers was compared with the distribution of answers given by the populations. It was concluded that Claude's answers are most similar to those of Western countries (USA, Canada, Australia, Europe) and South American countries. According to the paper, the results show *"potential for embedded biases in the models that systematically favor Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations"*. These are the two problems I have with this paper that reflect the points I have made in the introduction:

(1) Is the political behavior consistent? We do not know how the model would behave in different contexts. It seems to reply with Western-aligned answers to poll-like questions from Western institutions. But we simply do not know how far this setup generalizes. In fact, the paper shows that the model is *steerable*, and can generate answers aligned with different countries when asked to do so. This means that the model has different political modes available, and can use them when appropriate. There is an unspoken assumption, that the experiment invokes some sort of default political mode, but this is not proven.

(2) Are the results robust? Very little was done to check for algorithmic bias in the answers. There are some pretty important caveats in the data. Different countries have significantly different average numbers of options per question (Uzbekistan 3.8, Denmark 7.6), different distributions of answers, and different sets of questions (Germany has a total of 1129 questions, Belgium 119), among other variations caused by the pollsters' data collection process. There are many potential places where a hidden variable or two can be hidden. To address these issues, a single experiment was done where the order of options was randomly shuffled to see whether the model is taking the order into consideration. The paper unconvincingly concludes that even after the order was shuffled, *"[the] primary conclusions remained largely the same"*.

2.1 My experiments

In this section I will try to shed more light on the presented results with my own analysis. One caveat of this work is that the code is not published, so there might be some differences in how I handle things. Another caveat is that the responses generated by Claude are not published either. Only aggregated scores per country are available. This severely limits what we can do with the results.

Uniform Model. The numbers reported in the paper are difficult to interpret. Is the difference in the Jensen–Shannon distance[3] between the USA (0.68) and China (0.61) meaningful? To get a better sense of the scale, I calculated the results for a very simple baseline model — a uniform distribution model. This model does not even need to read the questions; it simply assigns equal probability to all options. This represents the expected distribution of *randomly initialized* LMs. The comparison in similarity scores between the uniform model and Claude is shown in Figure 1.

For the majority of countries, the uniform model outperforms Claude. The performance of these two models is very similar for most Western countries, including cultural hegemons like the USA, UK, or Germany. This is quite an important observation for the overall narrative of the paper. Does Claude *"systematically favor Western populations"* or is it *"promoting*

[2] This letter was heavily inspired by a previously published blog. Experimental code is available here.

[3] Jensen-Shannon distance is the measure of *alignment* used in the paper. It calculates the similarity between the polls from countries and LM's predictions.

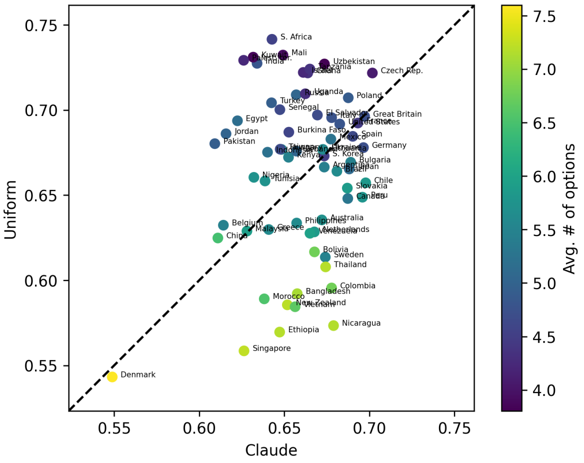


Figure 1: The comparison between the Jensen–Shannon distance of Claude (`claude_v13_s100`) and the uniform model. The average similarity is 0.659 for Claude and 0.664 for the uniform model. The uniform model wins in 53.8% of the countries.

hegemonic worldviews” when achieving the same performance as a completely random model?

Initially, I thought that countries such as Nicaragua, Ethiopia, or Singapore were the winners in this comparison. Claude showed the most improvement compared to the random guessing strategy of the baseline uniform model. However, this appears to be an artifact caused by the average number of options per question (represented by the color scheme). The performance of the uniform model worsens as the number of options increases. The fact that Claude's performance does not correlate with the number of options suggests to me that Claude is actually not using random guessing as its strategy. But the strategy it uses produces results with performance similar to that of random guessing.

Helpful. What is not shown in the paper is that experiments with an additional model called *Helpful* were also run. Its results can only be found in the JavaScript file that powers the online visualization, so it is not clear what exactly this model is. The Jensen-Shannon distance of various models is shown in Figure 2. *Helpful* significantly outperforms both Claude and the uniform model. It is better in all countries. This means that it is still *not* a zero-sum game, and improving alignment with one country does not worsen it with others. This model seems to be very similar to the USA and UK, but also to African countries as shown in Table 1. On the other hand, some Western countries are in the bottom 10. Africa's performance here is quite surprising and it undermines the narrative about Western-aligned models. Either the supposedly Western-centric nature of the data were somehow mitigated, or this is just some sort of a noise artifact. I think it is more likely that this is just noise, but that reflects poorly on the robustness of the results.

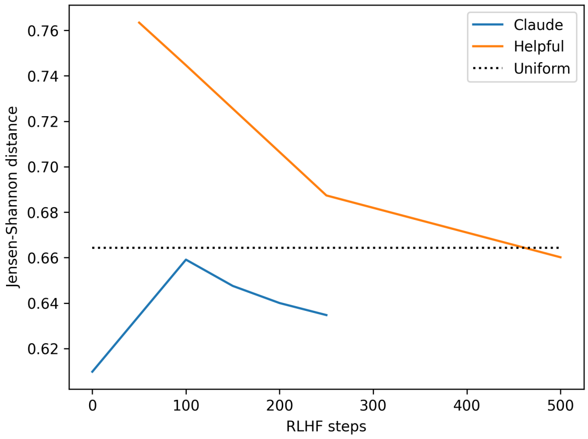


Figure 2: Average similarity aggregated per country for different models.

Top 10		Bottom 10	
United States	0.81	South Korea	0.74
United Kingdom	0.80	Pakistan	0.74
South Africa	0.80	Greece	0.74
Ethiopia	0.80	China	0.74
Mali	0.79	Sweden	0.74
Kenya	0.79	Thailand	0.74
Bolivia	0.79	Taiwan	0.74
Ghana	0.79	New Zealand	0.74
Nigeria	0.79	Belgium	0.69
Chile	0.79	Denmark	0.60

Table 1: The average similarity of the opinions aggregated per country.

Interpretation. Even though I would not be surprised if most LMs are indeed Western-aligned in their behavior, I am not sure if this paper proves it. Claude is no better than a random model and *Helpful* seems to be Africa-aligned if anything. **The results of the self-report study do not seem to be robust.** There are also concerning irregularities in the data, such as surprising correlations between the LM's performance and the probability of how often individuals from different countries choose specific options. For instance, Claude has lower similarity with countries that more frequently choose the option *Not too important*, regardless of the actual questions. Other strong correlations are shown in Table 2.

Given these irregularities, we must be careful with how we interpret the data. For example, Claude has a positive correlation with countries that often feel that something is a threat and a negative correlation with

Option wording	# Questions	Pearson's *r*
Not too important	44	-0.62
Somewhat favorable	54	0.61
Not a threat	36	-0.58
Major threat	36	0.56
Mostly disagree	44	0.52

Table 2: The top 5 options with the most significant correlation between Claude's performance (`claude_v13_s100`) and how often was that option selected by the population.

countries that do not feel threatened that much. There are multiple explanations for this behavior. (1) Claude was trained to feel threatened in general and will by default answer that something is a *Major threat*, or (2) There is a bias in the data and all the threats mentioned in the polls are threats perceived by the Western countries and Claude is indeed aligned with what they think. Both options are problematic. In the first case, we are not measuring a political opinion at all. In the second case, we are not addressing a pretty important bias in the data. Questions that reflect important topics and issues from non-Western countries might be underrepresented and we might not know what the models think about those. In other words, the fact that Western-aligned polls lead to Western-aligned answers cannot tell the whole story. **Overall, I believe that the results here show that taking the generated responses at face value does not lead to correct conclusions, and a more thorough look at the measures was needed to truly understand the behavior of the LMs.**

3 Feng et al. (2023)

The main idea of this paper is to measure the political leaning of LMs with the popular *Political Compass* online quiz. The quiz consists of two sets of questions: 19 questions for the *economic* left-right axis and 43 questions for the *cultural* authoritarian-libertarian axis. Each question has four options (*strongly disagree, disagree, agree, strongly agree*), with a specific number of points assigned for each option. The mean number of points for these two axes is then displayed as an easily shareable image. There are three main issues I have with this paper.

Validity. I find the use of this tool to be a shaky idea right out of the gate. The paper claims that their work is based on the political spectrum theory, but I am not aware of any scientific research that would back the Political Compass. To my knowledge, it really is merely a popular internet quiz with a rather arbitrary methodology based on the authors' intuition. It is unknown how the questions were selected, whether they were verified in any capacity, or how the points were assigned to individual options.

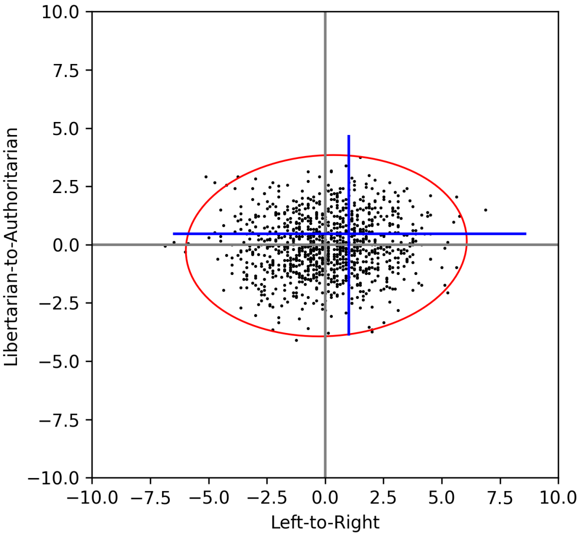


Figure 3: The Political Compass scores achieved by 1,000 random samples. The red circle shows the 3σ confidence ellipse. The blue cross shows the 3σ CIs for the two axes for a randomly selected sample.

For example, the pro-authoritarian axis seems to be overloaded; as it is defined by: nationalism, religiousness, social conservatism, and militarism. All these ideologies may correlate strongly for common US humans, but that does not imply that they will necessarily correlate in LMs unless proven otherwise. **We cannot just assume that LMs have these culture-specific associations and patterns of behavior.** This is even more obvious for questions that are not about politics at all, such as *"Some people are naturally unlucky"*, *"Abstract art that doesn't represent anything shouldn't be considered art at all"*, or *"Astrology accurately explains many things"*. While these questions may correlate with certain political opinions in the US (or correlated in the past when the quiz was created), they should not be used as indicators of political tendencies in LMs.

Statistical power. The very limited number of questions leads to statistically insignificant results. Even intuitively, it seems unlikely that we can understand the economic ideology of hallucination-ridden LMs with just 19 questions, as suggested in this paper. For comparison, I sampled a random model 1,000 times. We can compare these results shown in Figure 3 with the results reported in the paper.

There are two important observations here: (1) The confidence intervals for the individual samples are huge and they often contain most of the other samples and all four political quadrants. Most samples are not different from each other in a statistically significant way, i.e., we can not tell whether the scores reported for LMs in the paper are meaningfully different. (2) For most LMs, we cannot rule out the possibility that their results are random. The only exception is the cultural axis for some of the LMs (e.g., GPT-J with a score of more than 5). Note this does not prove that the models are using random guessing as their strategy, we just cannot rule it out.

Downstream evaluation. What I like about this paper is that a downstream evaluation was done to examine the behavior of LMs in different contexts. LMs were trained with politically biased data (e.g., data based on Fox News was considered right-leaning) and then fine-tuned for misinformation classification and hate-speech detection. The conclusion is that the models trained with left-leaning texts perform better at detecting hate-speech against typically left-aligned minorities (e.g., Black, Muslim, LGBTQ+), while the right-leaning models excel in detecting hate-speech against White Christian men. Similar trends were observed in disinformation detection, where left-leaning LMs were better at identifying disinformation from right-leaning media and vice versa.

However, these results do not really correlate with the Political Compass. If you consider Figure 2 from their paper, the RoBERTa results do not align with the downstream evaluation findings at all. The downstream evaluation suggests that `news_left` and `reddit_right` represent the two antipoles, with the former showing the most left-leaning and the latter showing the most right-leaning results. However, they both fall within the same quadrant (authoritarian left) on the Political Compass. **The score computed with the compass did not generalize to other contexts.** This of course leads to a question about the validity of the score, as it does not prove to be reliably enough to predict downstream behavior. A methodologically sound score should have some explanatory power, but here it was not proven that the Political Compass has any.

4 Nadeem et al. (2021)

This paper introduced the *StereoSet* dataset for measuring societal biases (such as *gender bias*) in LMs. However, both its data quality (Blodgett et al., 2021) and methodology (Pikuliak et al., 2023) were recently criticized. The flaws identified in the latter paper are connected to the faulty assumptions about using self-report studies, so they can serve as a good illustrative example for the purposes of this letter. I will reuse their findings and recontextualize them here.

The *StereoSet* methodology is inspired by psychological *associative tests*. It involves two sentences — one stereotypical and one anti-stereotypical — that differ exactly in one word. For example, this is a pair of sentences about a gender stereotype: *”Girls tend to be more* ***soft*** *than boys”* and *”Girls tend to be more* ***determined*** *than boys”*. We mask the position of the keyword and ask an LM to fill it in. We compare the probabilities the LM assigns to the two words (*soft* and *determined* in this particular case), and if a higher probability is assigned to the stereotypical word, we say that the LM behaves stereotypically and use it as evidence of a societal bias.

A test like this intuitively makes sense for humans. Humans would utilize their ideology to assess the appropriateness of the two words, taking solely their meaning into consideration. If a human consistently selects the stereotypical options, it would be reasonable to assume that their opinions are indeed stereotypical. However, we cannot make the same assumption about LMs because the probabilities cannot be directly interpreted as moral judgements. This statement can be illustrated with the two following experiments.

(1) LMs tend to select more frequent words. Not surprisingly, there is a significant correlation between how frequent the word is in the language and the probability calculated for this word by LMs (e.g., Pearson's r of 0.39 for gender bias with `roberta_base`, see Figure 4). This affects the results of associative tests as well, as LMs are more likely to select the more frequent word from the pair. Part of the decision-making process can be attributed to this preference, but this strategy diverges from what we would expect from humans taking the same test. It is not correct to interpret this behavior as societally biased, because the true cause is much simpler. Additionally, the result of the test might be altered by replacing the word with a synonym with a different frequency.

(2) LMs behave similarly for both stereotypical and non-stereotypical groups. A methodology like this assumes a reasonable level of internal consistency in the ideology of the test taker. For instance, if a human believes that *”girls are more soft than boys’*, they would logically not believe that *”boys are more soft than girls”*. Are LMs consistent like that? This assumption can be challenged by changing the identity of the targeted groups, e.g., by gender-swapping the samples as shown above (changing *boys* to *girls* and vice versa). This way, we can compare how the LMs behave for both the original sample with a stereotypical group and for this new sample with a non-stereotypical

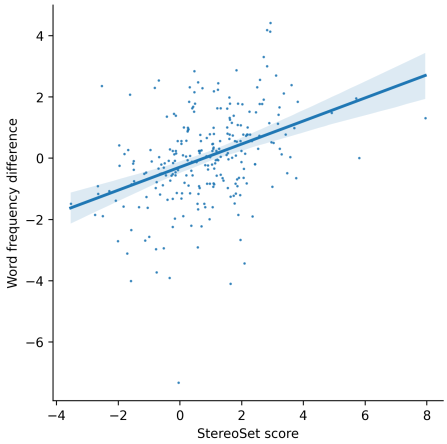


Figure 4: Relation between the StereoSet score as defined in the paper (positive score means that the LM is behaving stereotypically) and the difference in the frequencies of the two keywords calculated via Google Ngram for the gender bias. Each point is one sample. `roberta_base` was used as the LM.

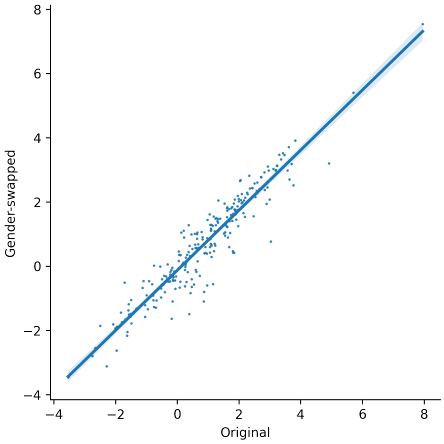


Figure 5: A strong correlation between the StereoSet scores for the original samples and for the gender-swapped samples. Results calculated for `roberta_base`.

group. Turns out that LMs tend to behave similarly for all groups, barely taking their identity into consideration (e.g., Pearson's r of 0.95 for gender bias with `roberta_base`, see Figure 5). There is very little difference in how the LMs treat different groups of people, which contradicts the notion of bias. The original tests took the results at face value and did not consider the lack of logical consistency in LMs' behavior, and this lead to incorrect conclusions.

Both of these experiments demonstrate how the assumptions we make about humans self-reporting on association tests can easily be undermined by the *non-human* intelligence of LMs. Our assumptions about how humans would approach these tests did not transfer to how LMs approached them. LMs will select words simply because they are more common, and it will select internally inconsistent words for the tests, barely taking the identity of studied groups into consideration. **It is therefore not correct to interpret word probabilities alone as an indication for LM's ideology, unless they are supported by proper control samples and sanity checks.**

5 Conclusions

I think it is safe to assume that LMs have various forms of political, psychological, societal, and other types of behavior baked in within. Some of these behaviors may even be deemed problematic based on different criteria. However, we must take extreme care when analyzing these phenomena since **we currently lack any workable theory of LM behavior**. Using self-report studies originally designed to study human intelligence is tricky, as highlighted in this letter with various failure modes found in the papers. Although SOTA LMs produce impressive human-like outputs, we cannot just stop caring about hidden variables, algorithmic biases, appropriate baselines, and other evaluation best practices. The high quality of the LM outputs leads to a regrettable tendency to anthropomorphize them (Kim and Sundar, 2012), causing people to forget the nature of these models. Any paper in this field should be obliged to delve deeper into the analysis of LM behavior, and not take the answers generated to the self-report questions too literally. Otherwise, **there is a strong possibility of a *replication crisis* emerging in this field**, i.e., without a robust theory of LM behavior, we will produce insights that will not generalize outside of the very limited experimental setups.

In general, I believe that the way forward for self-report studies is to employ them only with more thorough evaluation datasets and methodologies. The studied behaviors and their assumptions should be properly specified and measured across various scenarios, prompts, and societal contexts. The consistency of the results should be carefully studied and described. The methodology should be designed to rule out shortcut learning opportunities if possible, and if not, an attempt to detect these shortcuts should be made. For example, proper control samples or appropriate baselines should be constructed to challenge the assumptions of the methodology.

References

Banyard, Philip, Andrew Grayson, and MT Orne. 1996. Demand characteristics. *Introducing psychological research: Sixty studies that shape psychology*, pages 395–401.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Blodgett, Su Lin, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Durmus, Esin, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Feng, Shangbin, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Hartmann, Jochen, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Hershcovich, Daniel, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Holtzman, Ari, Peter West, and Luke Zettlemoyer. 2023. Generative models as a complex systems science: How can we make sense of large language model behavior? *arXiv preprint arXiv:2308.00189*.

Huang, Jen-tse, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Kim, Youjeong and S Shyam Sundar. 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, 28(1):241–250.

Ma, Pingchuan, Zongjie Li, Ao Sun, and Shuai Wang. 2023. " oops, did i just say that?" testing and repairing unethical suggestions of large language models with suggest-critique-reflect process. *arXiv preprint arXiv:2305.02626*.

Nadeem, Moin, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Narayanan, Arvind and Sayash Kapoor. 2023. Evaluating LLMs is a minefield. `https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield/`. Accessed: 2023-12-09.

Nass, Clifford, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78.

Pikuliak, Matúš, Ivana Beňová, and Viktor Bachratý. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3648–3665, Dubrovnik, Croatia. Association for Computational Linguistics.

Rudiak-Gould, Peter. 2010. Being marshallese and christian: A case of multiple identities and contradictory beliefs. *Culture and Religion*, 11(1):69–87.

Rutinowski, Jérôme, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of chatgpt. *arXiv preprint arXiv:2304.07333*.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Schuman, Howard and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Weizenbaum, Joseph. 1976. Computer power and human reason: From judgment to calculation.

Wood, Michael J, Karen M Douglas, and Robbie M Sutton. 2012. Dead and alive: Beliefs in contradictory conspiracy theories. *Social psychological and personality science*, 3(6):767–773.

Generation and Evaluation of Multiple-choice Reading Comprehension Questions for Swedish

Dmytro Kalpakchi, KTH Royal Institute of Technology, Stockholm, Sweden dmytroka@kth.se
Johan Boye, KTH Royal Institute of Technology, Stockholm, Sweden jboye@kth.se

Abstract Multiple-choice questions (MCQs) provide a widely used means of assessing reading comprehension. The automatic generation of such MCQs is a challenging language-technological problem that also has interesting educational applications. This article presents several methods for automatically producing reading comprehension questions MCQs from Swedish text. Unlike previous approaches, we construct models to generate the whole MCQ in one go, rather than using a pipeline architecture. Furthermore, we propose a two-stage method for evaluating the quality of the generated MCQs, first evaluating on carefully designed single-sentence texts, and then on texts from the SFI national exams. An extensive evaluation of the MCQ-generating capabilities of 12 different models, using this two-stage scheme, reveals that GPT-based models surpass smaller models that have been fine-tuned using small-scale datasets on this specific problem.

1 Introduction

In several educational stages, multiple-choice questions (MCQs) provide a widely used means of assessing reading comprehension (OECD, 2021). Tests that consist of MCQs have the very appealing property of allowing swift, automatic, and objective grading. However, creating such tests is far from being swift or automatic, but rather is time-consuming and requires a great deal of expertise (Haladyna, 2004). In this work, we analyze to what extent the creation of MCQ tests for reading comprehension in Swedish could be automated using publicly available language models (both closed-source models via public APIs, and open-source ones).

The focus of this article is on MCQs aimed at assessing reading comprehension of second-language learners of Swedish, specifically aimed at the Swedish for Immigrants courses (SFI). Our contributions[1] are:

- we propose a number of methods that can generate a number of distinct reading comprehension MCQs from a given text;
- we propose a two-stage method for evaluating the quality of the generated MCQs, first evaluating on carefully designed single-sentence texts, and then on a small corpus of texts[2] from the SFI national exams (for the D-level course);
- we compare our proposed methods with the state-of-the-art GPT-3 and ChatGPT models, as well as some baselines.

An MCQ consists of a question proper (the *stem*), the correct answer (the *key*), and a number of wrong but plausible options (the *distractors*). We will refer to the key and distractors together as *alternatives* (ALT). Contrary to prior work (Majumder and Saha, 2015; Araki et al., 2016; Guo et al., 2016; Zhou et al., 2020; Kalpakchi and Boye, 2021), we do NOT split the problem of generating stem and key from the problem of generating distractors. Instead, we aim to generate the whole "package" at once and be able to offer more than one MCQ per text. We assume that the texts are already given, and that they are on the appropriate level for testing reading comprehension, e.g., they are of the appropriate length, split into paragraphs, use the vocabulary of the appropriate complexity for the second-language learners, etc. In this article we do NOT conduct any assessment on how appropriate the given texts are for testing reading comprehension. The interested reader is referred to (OECD, 2019, 2021) for more discussion on this matter.

In the aforementioned prior work, researchers have tried a two-stage approach to MCQ generation, first generating a stem-key pair using one method and then generating the distractors using another method. Such approaches have a number of advantages, e.g., the key can be extracted directly from the text and thus guar-

[1] The source code and the Plugga corpus will be freely available upon the publicaiton of the article.

[2] We call this corpus *Plugga* and make it available online

Licensed under CC BY 4.0, pages 86–105, DOI 10.3384/nejlt.2000-1533.2024.4886

anteed to be correct, or the stem/key formulation can be edited to (hopefully) get higher quality distractors. Nevertheless, generating the whole "package" at once, as we attempt to do in this article, has also its advantages. Our motivation is twofold. The first reason is that a stem-key pair produced at stage 1 might not necessarily allow for good distractors to be produced at stage 2. The hope is that when generating the whole MCQ in one go, the model will learn to only generate stems that have reasonable alternatives. The second reason is speed of execution, meaning that it is more resource-efficient to run one model and directly obtain the entire MCQ, instead of running several models that generate each part of the MCQ separately.

2 Related work

Automatic question generation from text has been studied before, mainly for the English language. Results obtained up to 2020 are summarized in the survey article by Zhang et al. (2021). However, very little work has been done on generating whole MCQs (rather than just the questions), although some researchers have focused on generating distractors separately using large language models (Qiu et al., 2020; Zhou et al., 2020; Offerijns et al., 2020; Chung et al., 2020; Kalpakchi and Boye, 2021).

To the best of our knowledge, there have been no prior attempts on generating *the whole reading comprehension MCQs* for Swedish using fully open-source and free-to-use models. The only prior attempt in this direction was by Kalpakchi and Boye (2023a), where they generated MCQs using OpenAI's GPT-3 (Brown et al., 2020), which is neither open-source nor free to use.

However, there have been attempts at generating parts of MCQs for Swedish. Kalpakchi and Boye (2022a) generated stems and keys separately using a data-driven rule inductor based on Universal Dependencies for creating templates for question-answer pairs and then attempting to apply them to single sentences during the generation process.

Kalpakchi and Boye (2021) experimented with a method based on the Swedish BERT (Malmsten et al., 2020) for generating distractors given the text, the stem, and the key.

3 Data

3.1 Training

In this work we experimented with two previously released datasets of Swedish MCQs for reading comprehension. The first one, SweQUAD-MC (Kalpakchi and Boye, 2021), contains MCQs on texts scraped from the websites of Swedish public authorities. Following the definition of OECD (2019), all texts have a *continuous* format, which means they have no internal structure beyond being organized into sentences and paragraphs. The MCQs were created by paid linguistics students and required both the key and distractors either to appear in the text directly, or be a grammatical reformulation of a phrase present in the text. Most MCQs in SweQUAD-MC contain three alternatives, but some include more (up to six alternatives). Nevertheless, it is guaranteed that exactly one of the presented alternatives is correct.

The second dataset, Quasi (Kalpakchi and Boye, 2023a), consists of 90 texts collected from the SFI[3] national examinations, along with 317 MCQs synthetically generated by GPT-3 and manually curated. The texts in Quasi are of different genres, e.g., news articles, ads, e-mails, blog posts, etc. The majority of the texts are either partially or fully *non-continuous* which, by the definition of OECD (2019), means that they have some internal structure that helps (or is necessary for) understanding their content. For instance, e-mails contain the addresses of the sender and receiver in certain places, recruitment ads feature the employer company name and contact details, and posts contain the name, and possibly also the age, of the author. All MCQs in Quasi contain four alternatives, of which exactly one is correct.

3.2 Evaluation

Evaluation sets for reading comprehension questions in Swedish are scarce. SweQUAD-MC does have a validation and test set, but the texts are not verified to be suitable for testing reading comprehension (e.g., some of them might use too complex language, especially for 2nd language learners). The texts in Quasi are taken from national SFI examinations and are thus suitable for assessing reading comprehension. However, Quasi provides only a small set of MCQs, which is impractical to split further into training and test sets.

It is also worth noting that there are currently no reliable methods to automatically evaluate the quality of MCQs. Evaluation methods like BLEU, ROUGE, and METEOR, which are based on word overlap between the generated result and the "gold" MCQ in a test set, will not give much information due to the open-ended nature of the task – from any non-trivial text, a very large number of MCQs are possible. Hence, there is no real benefit in having a test set of MCQs. Instead, one should have a test set of texts suitable for reading comprehension, and with a degree of variation in multiple aspects, such as length, genre, formatting, etc.

In this work, we have adopted the latter approach

[3] Swedish for Immigrants, the national Swedish course curriculum for 2nd language learners.

and have collected a corpus of 10 texts for evaluating reading comprehension in Swedish, a corpus which we will refer to as *Plugga*. Similarly to designers of Quasi, we took materials from the previous national exams for SFI (but made sure that there is *no overlap* between Quasi and Plugga) by running the Tesseract OCR engine[4], and manually correcting the outputs. The sources and genres of texts in Plugga are distributed as follows:

- 2 newspaper articles, shortened and simplified by the SFI test constructors;
- 1 shortened and simplified yearly report from the public authority (Statistics Sweden, SCB);
- 1 short text with tips when to ring the emergency telephone number 112 from SOS Alarm (the company running 112);
- 2 compiled short answers to a given question by multiple people;
- 2 e-mails;
- 1 short forum thread discussing a given issue;
- 1 detailed program to an event.

The first three sources are continuous texts, divided into paragraphs, whereas the last four sources are either fully or partially non-continuous.

4 Method

In this work, we have experimented with fine-tuning two publicly available large language models: Swedish BERT (Malmsten et al., 2020), and SweCTRL-Mini (Kalpakchi and Boye, 2023c). We compared the fine-tuned models to two baselines described in Section 4.4, as well as GPT-3 (Brown et al., 2020), specifically *text-davinci-003*, and ChatGPT[5], specifically *gpt-3.5-turbo-0301*. We did not use GPT-4[6] in this work, because at the time of writing, access to its API is limited for the general public. The same goes for GPT-SW3[7].

4.1 Models based on KB/BERT

Swedish BERT, later referred to as *KB/BERT*, is a discriminative model that has been previously used by Kalpakchi and Boye (2021) for generating distractors with relative success. In this work, we also use KB/BERT, but attempt to generate *whole MCQs* and not only distractors. Following Kalpakchi and Boye (2021), we frame the problem as an auto-regressive generation in two different ways: left-to-right and arbitrary-order[8]. The training procedure for both cases is summarized in Table 1. In both cases, each MCQ (here with 2 distractors) is represented as follows (later referred to as an *MCQ sequence*):

```
T [SEP] Q [SEP] A [SEP] D1 [SEP] D2
```

where `Q` denotes all the tokens of the stem (the question proper), `A` the words in the key (the correct answer), and so on, and `[SEP]` is the special separator token in BERT. Each `[SEP]`-separated part of the MCQ sequence except `T` will be referred to as an *MCQ sequence item*. For the sake of brevity, we will only use two distractors `D1` and `D2` in the examples. In general, we will use D to denote the set of distractors.

For the **left-to-right variant** (LRV), both training and generation is designed to proceed from left to right both when producing the whole MCQ sequence, and when generating each MCQ sequence item. **At training time**, each MCQ from the training data is represented as an MCQ sequence, which is then converted into multiple datapoints. This conversion is obtained by building the MCQ sequence one token at a time, masking the last token, and attempting to predict it. An example of such conversion is given in the top subtable of Table 1. In row 1, we started building the MCQ sequence, which consists of a single token. This token gets masked, and the task is to predict the correct token `Q1` from the Target column. Then, we add the next token (row 2) and mask the last token in the partial MCQ sequence, and again attempt to predict that token (`Q2` in this case), and so on. When we finished generating the stem (row 4), we add the `[SEP]` token, which now becomes the last token of the sequence and hence is also masked, requiring the model to be able to also predict `[SEP]` tokens correctly for learning to finish each sequence item. In this manner, each MCQ is converted into $|\texttt{Q}|+|\texttt{A}|+\sum_{\texttt{DX}\in D}|\texttt{DX}|+|D|+2$ training datapoints. **At generation time** the process is similar to the training time, but without knowing the target tokens. Specifically, we input the text `T`, followed by a `[SEP]` token, and append a `[MASK]` token at the end. Then we unmask the `[MASK]` token, by sampling from the provided distribution, and append a new `[MASK]` token. This process is repeated until we have generated $|D|+2$ `[SEP]` tokens, in which situation we assume that the stem, the key, and all distractors have been generated. We have enforced a hard limit of 30 tokens for the stem and 20 tokens for each alternative.

For the **arbitrary-order variant** (AOV-A), both training and generation is designed to proceed in an

[4] Freely available at https://github.com/tesseract-ocr/tesseract

[5] https://openai.com/blog/chatgpt

[6] https://openai.com/gpt-4

[7] https://www.ai.se/en/node/81535/gpt-sw3

[8] Referred to as the "u-PMLM variant" in the original article.

#	Input for left-to-right KB/BERT variant (LRV)	Target
1	[M]	Q1
2	Q1 [M]	Q2
3	Q1 Q2 [M]	Q3
	...	
4	Q1 Q2 ... QK [M]	[S]
5	Q1 ... QK [SEP] [M]	A1
6	Q1 ... QK [S] A1 [M]	A2
7	Q1 ... QK [S] A1 A2 [M]	[S]
8	Q1 ... QK [S] A1 A2 [S] [M]	D11
9	Q1 ... QK [S] A1 A2 [S] D11 [M]	D12
10	Q1 ... QK [S] A1 A2 [S] D11 D12 [M]	[S]
11	Q1 ... QK [S] A1 A2 [S] D11 D12 [S] [M]	D21
12	Q1 ... QK [S] A1 A2 [S] D11 D12 [S] D21 [M]	D22
13	Q1 ... QK [S] A1 A2 [S] D11 D12 [S] D21 D22 [M]	D23
14	Q1 ... QK [S] A1 A2 [S] D11 D12 [S] D21 D22 D23 [M]	[S]

#	Input for arbitrary-order KB/BERT variant (AOV-A)	Target(s)
1	Q1 [M] Q3 ... QK	Q2
2	[M] Q2 [M] ... QK	Q1, Q3
3	[M] [M] Q3 ... [M]	Q1, Q2, QK
	...	
4	Q1 ... QK [S] [M] A2	A1
5	Q1 ... QK [S] [M] [M]	A1, A2
6	Q1 ... QK [S] A1 [M]	A1
7	Q1 ... QK [S] A1 A2 [S] [M] D12	D11
8	Q1 ... QK [S] A1 A2 [S] D11 [M]	D12
9	Q1 ... QK [S] A1 A2 [S] [M] D12	D11
10	Q1 ... QK [S] A1 A2 [S] D11 D12 [S] D21 D22 [M]	D23
11	Q1 ... QK [S] A1 A2 [S] D11 D12 [S] [M] D22 [M]	D21, D23
12	Q1 ... QK [S] A1 A2 [S] D11 D12 [S] D21 [M] D23	D22

#	Input for arbitrary-order-all-at-once KB/BERT variant (AOV-B)	Target(s)
1	Q1 [M] Q3 Q4 [B] [S] [M] A2 [B] [S] D11 D12 [B] [S] D21 D22 D23	Q2, A1
2	Q1 Q2 [M] Q4 [M] [S] A1 [M] [B] [S] D11 [M] [B] [S] D21 D22 D23	Q3, [B], A2, D12
3	[M] Q2 Q3 [M] [B] [S] [M] [M] [B] [S] D11 D12 [B] [S] D21 D22 D23	Q1, Q4, A1, A2
4	Q1 Q2 Q3 Q4 [M] [S] A1 A2 [M] [S] D11 D12 [M] [S] D21 D22 D23	[B], [B], [B]
5	[M] Q2 Q3 Q4 [B] [S] [M] A2 [B] [S] D11 [M] [B] [S] D21 D22 D23	Q1, A1, D12
6	Q1 Q2 Q3 Q4 [B] [S] A1 [M] [B] [S] D11 D12 [B] [S] D21 D22 D23	A2
7	Q1 [M] Q4 [M] [S] A1 A2 [B] [S] D11 D12 [B] [S] D21 D22 D23	Q2, [B]
8	Q1 Q2 Q3 Q4 [B] [S] A1 A2 [B] [S] D11 [M] [B] [S] [M] D22 D23	D12, D21
9	Q1 Q2 Q3 Q4 [B] [S] A1 A2 [M] [S] D11 D12 [B] [S] D21 D22 D23	[B]

Table 1: Example datapoints extracted from *one* MCQ for training the model capable of left-to-right (top table) or arbitrary-order generation (bottom table). **Observe that all inputs are also prefixed with a string** `[CLS] T [SEP]`, **which is omitted in this table for the sake of brevity**. `[M]` and `[S]` denote BERT's `[MASK]` and `[SEP]` tokens respectively. `[B]` denotes a special padding token `[BLANK]` introduced by us. In the example for the AOV-B variant, the question was padded to the maximum of 5 tokens, and each alternative was padded to the maximum of 3 tokens.

arbitrary order only when generating each MCQ sequence item, whereas the whole MCQ sequence is still produced from left to right. **At training time**, in contrast to LRV, we pad the stem and each alternative with a specially introduced token `[BLANK]` to obtain a fixed width of 30 tokens for the stem and 20 tokens for each alternative (the same as the hard limits for LRV). Again we convert each MCQ into multiple datapoints, but in a different way. We start with the leftmost MCQ sequence item, and *corrupt* it K times, as follows: We first draw a masking probability r from the uniform distribution, and attempt to mask each token of the MCQ sequence item with this probability. For instance, in rows 1–3 of the middle sub-table from Table 1, we deal with corrupting first MCQ item, the stem (question proper). Note that the masking in each row corresponds to a *different* value of r (and there are K such values in total). Once we have corrupted one MCQ sequence item K times, we append its non-corrupted version to the partial sequence, and proceed with corrupting the next sequence item, the key, while keeping the preceding MCQ sequence items non-corrupted. We proceed in the same manner until the whole MCQ sequence has been processed. Following (Kalpakchi and Boye, 2021), and contrary to LRV, we don't train the model to unmask the `[SEP]` token. **At generation time**, we provide a fixed number of `[MASK]` tokens (30 for the stem, and 20 for each alternative). Following Kalpakchi and Boye (2021), we unmask the token at the position where the model is most confident. However, rather than selecting the token with the maximum probability, we sample, in order to be able to generate more than one MCQ per text.

Additionally, we introduce another arbitrary-order setup, where we perform exactly the same procedure as for AOV-A, but on all MCQ sequence elements at once (as demonstrated by the bottom sub-table of Table 1). We refer to this setup as AOV-B. **At generation time**, we add $30 + 20 \cdot (|D| + 1)$ `[MASK]` tokens separated by `[SEP]` tokens. Apart from this, the unmasking procedure is the same as for AOV-A.

Because of the elaborate unmasking scheme, the generation phase of AOV-A and AOV-B takes somewhat longer time per token than the LRV setup.

4.2 Models based on SweCTRL-Mini

SweCTRL-Mini is a generative model capable of generating texts one token at a time, left to right. We fine-tune it similarly to left-to-right generation for KB/BERT, except that we don't use `[MASK]` and `[SEP]` tokens, as they are BERT-specific. Instead, we introduce a new *control code pair* consisting of the *opening control code* `:mcq:`, and its corresponding *ending control code* `:mcq:$`. The key is always the first of the four alternatives and is prefixed by `a)`, whereas all distractors are prefixed by the letters after "a", i.e., `b)`, `c)`, etc. Hence the structure of a datapoint for fine-tuning SweCTRL-Mini would look as follows:

```
T :mcq:  Q a) A b) D1 c) D2 :mcq$
```

We then train the model to predict one token at a time, left to right, for all tokens except those in `T`. **At generation time**, the MCQs were sampled we append `:mcq:` after the text `T` and sample the new tokens left to right using the nucleus sampling with the threshold $p = 0.9$ until reaching `:mcq:$`.

4.3 Models based on GPT

We used both GPT-3 and ChatGPT in a zero-shot manner. Inspired by Kalpakchi and Boye (2023a), we input the following prompt to both models, in **Swedish**:

> Skriv N_q^T läsförståelsefrågor med alternativ (a, b, c, d, o.s.v.) och ge varje fråga en unik nummer (1, 2, 3, o.s.v.). Första alternativet (a) ska alltid vara rätt, medan de andra alternativen (b, c, d, o.s.v.) ska vara felaktiga, men troliga. Alla frågor måste kunna besvaras av den följande texten.

To aid the readers not speaking Swedish, we provide the prompt's English *translation* below:

> Write N_q^T reading comprehension questions with alternatives (a, b, c, d, and so on) and give each question a unique number (1, 2, 3, and so on). The first alternative (a) should be always correct, while the other alternatives (b, c, d, and so on) should be wrong, but plausible. All questions must be answerable by the following text.

However, we stress again that the prompt was fed directly *in Swedish*. We used nucleus sampling with the nucleus threshold $p = 0.9$, and limited the maximum number of tokens to 2048 to accommodate the larger texts in Plugga.

4.4 Baselines

4.4.1 Baselines using Universal Dependencies

For this baseline, we generate stems and keys first using Quinductor (Kalpakchi and Boye, 2023b, 2022a) and then use the extractive distractor generation baseline proposed by Kalpakchi and Boye (2021). For both of these, we have used the provided official implementations, available on GitHub. All of these methods rely on Universal Dependencies (Nivre et al., 2020) and require the following resources:

- a pre-trained dependency parser compliant with Universal Dependencies;

- a dataset of texts and question-answer pairs (QA-pairs) for automatically inducing the Quinductor templates;
- a corpus of texts to extract the distractors from (referred to as *DIS-corpus*).

As a minimal example of a Quinductor template, consider the sentence “Tim plays basketball”, with an associated question “What does Tim play?”. The parser would conclude that the sentence “Tim plays basketball’ consists of a verb (the root node `r`), the subject (`r.nsubj`), and an object (`r.obj`). The question “What does Tim play?” can then be described as “What does `[r.nsubj] [r.lemma]`?”. This template can then be used to generate a question from a new, previously unseen sentence with a parallel grammatical structure (e.g. “Sue likes spaghetti” yielding “What does Sue like?”). Note that such templates are induced (and can be used) only on single sentences in the text.

When all the resources are in place and the Quinductor templates have been induced, the generation of an MCQ with K alternatives based on the previously unseen text `T'` should proceed as follows:

1. Using the Quinductor method, generate and rank the QA-pairs using previously induced templates.
2. Select the desired number $N_q^{T'}$ of highest-ranked QA-pairs
3. For each QA-pair, extract distractors from the DIS-corpus by searching for the first $K-1$ syntactic structures similar to that of the key.

4.4.2 Zero-shot SweCTRL-Mini

Recall that SweCTRL-Mini is a generative model, in contrast to KB/BERT. Hence, it is possible that it could be able to produce MCQs (fully or partially) in a zero-shot manner. In this work we experiment with the following setup (later referred to as simply *Zero-shot*):

- Input the text `T` (for longer texts we follow the procedure outlined in Section 5).
- First generate a stem by using the prompt “`T Fråga:`” and keep generating until a “?” symbol is produced (separately or as part of another token).
- Then, use the generated stem and attempt to generate the key with the prompt “`T Fråga: Q Svar:`”. Proceed until generating a full stop (.).
- Finally, use the generated stem and key and attempt to generate three distractors with the prompt “`T Fråga: Q a) A b)`”. Terminate generation when reaching either the string “`e)`” or the string “Fråga”.

At all stages of the generation process, we imposed a hard limit of 30 tokens.

5 Experimental setup

In this work we fine-tuned models for the 4 proposed methods (KB/BERT LRV, KB/BERT AOV-A, KB/BERT AOV-B, and SweCTRL-Mini). Each of these models was trained on two datasets (SweQUAD-MC, and Quasi), resulting in $4 \cdot 2 = 8$ models.

Each fine-tuned model was trained for 10 epochs using the AdamW optimizer (Loshchilov and Hutter, 2019) with the default Huggingface training parameters: initial learning rate 5×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, without learning rate scheduling or weight decay. The gradients were clipped to the norm of 1. The training was conducted on a single NVIDIA 3090 GPU with 24GB of VRAM using a batch size of 8 for models based on SweCTRL-Mini, and a batch size of 16 for those based on KB/BERT. The checkpoints for each model were saved for each epoch, resulting in 10 checkpoints per fine-tuned model.

The major challenge with both KB/BERT and SweCTRL-Mini is their limited and rather small context window size (512 tokens for KB/BERT, and 256 tokens for SweCTRL-Mini). At all times the context window should accommodate both the text and the MCQ. To ensure that, we limited the number of text-related tokens to at most L_T tokens. The exact value of this limit was model-specific, namely $L_T = 441$ for KB/BERT LRV, $L_T = 384$ for KB/BERT AOV-A and AOV-B, and $L_T = 192$ for SweCTRL-Mini. However, if the MCQs in the training data could not be fit in the remainder of the context window, we automatically decreased these limits[9]. At all times we ensure that the basis for the correct answer from the text (the information that is provided by both datasets) is included in the context window.

For the UD-based baseline, we relied on the training set of SweQUAD-MC for generating both QA-pairs[10], and distractors[11]. For this work, we opted out of inducing such templates on Quasi, because most of the texts in Quasi are (partially) non-continuous texts. Since Quinductor was designed to work on single sentences from continuous texts, it is therefore unlikely that templates induced on Quasi will end up being generalizable (or will be induced at all).

The zero-shot SweCTRL-Mini baseline did not require any specific further training (by the nature of being zero-shot). This brings the total number of models to ten.

[9] For more information on this heuristic we refer to the source code accompanying the article.

[10] Using Quinductor templates provided by Kalpakchi and Boye (2022a) in the associated GitHub repository

[11] Using only raw texts from the training set of SweQUAD-MC

When evaluating, similarly to Kalpakchi and Boye (2023a), the longer the text T, the more MCQs we attempted to generate. More specifically, we requested N_q^T MCQs for each T using the following formula:

$$N_q^T = \left\lceil \frac{C_T}{\bar{W} \cdot \bar{C}} \right\rceil, \tag{1}$$

where C_T is the number of characters in T, $\bar{W} = 12.78$ ($\bar{C} = 4.81$) is the average number of words (characters) per sentence. These quantities were calculated as a weighted average across corpora from (Kalpakchi and Boye, 2023a, Table 1) with the weights being the relative sizes of the corpora in words.

6 Model selection

The goal for this section is to select the best checkpoint for each of the eight fine-tuned models. Since there are ten checkpoints per model, this step requires evaluating the quality of 80 checkpoints, which is prohibitively expensive to do using human evaluation. Instead we resort to using metrics that could be calculated automatically. Furthermore, since there is no reliable way to estimate how good the MCQs are using the automatic metrics, we aim at estimating how many *definitely* bad MCQs were generated by each checkpoint.

To define *definitely* bad MCQs we employed the following *badness metrics* (listed from the most to the least severe). In the list below, "MCQ%" means "percentage of MCQs", and "ALTs" means "alternatives" (the key and distractors together), whereas "ALT" means "an alternative" (either the key or any distractor). For all of the metrics below, *the lower, the better.*

1. *AltInStem.* MCQ% with any ALT being verbatim in the stem.
2. *AltAllSame.* MCQ% with all identical ALTs.
3. *StemTextRep.* MCQ% with the stem containing repetitive phrases contiguously (up to 10 tokens).
4. *AltAnyTextRep.* MCQ% with any ALT containing repetitive phrases contiguously (up to 10 tokens).
5. *AltAnySame.* MCQ% with ≥ 2 (but not all) identical ALTs.
6. *AltAnyEmpty.* MCQ% with ≥ 1 ALTs being an empty string[12].
7. *StemEmpty.* MCQ% with the stem being an empty string[12].
8. *StemNoQmark.* MCQ% with the stem not ending with a question mark.
9. *NoEndCode.* MCQ% where the generation was not finished with the appropriate control code `:mcq$:` (only for models based on SweCTRL-Mini).

We evaluated all checkpoints on the texts from the development set of SweQUAD-MC. The texts that are larger than L_T tokens are split into chunks of max L_T tokens. For each model, we have generated N_q^T MCQs, as calculated by Equation 1. Because generating MCQs using models based on KB/BERT is computationally heavy (and we need to generate MCQs for 80 checkpoints), we calculated the badness metrics only for the MCQs generated on the first 100 text chunks from the development set (when all the texts are sorted alphabetically).

Based on the badness metrics reported in Figure 1, we selected the checkpoint with the fewest and least severe errors for each model (recall that the introduced badness metrics are listed from the most to the least severe). Additionally, everything else being the same, we preferred earlier checkpoints to reduce the risk of overfitting (given that the training sets were quite small, especially for the SweCTRL-based models). The selected checkpoints per model (denoted by the number of training epochs) are reported in Table 2.

7 Human evaluation

For human evaluation, we compare eight best checkpoints selected in Section 6, two baseline models, and two GPT-based models, namely GPT-3 (*text-davinci-003*), and ChatGPT (*gpt-3.5-turbo-0301*).

In this section the evaluation begins with the following basic question for the set of produced MCQs per model per text:

Q0. Did the model produce the requested number N_q^T of MCQs?

Then each generated MCQ is evaluated separately on the following aspects:

Q1. Are the question (stem) and all alternatives grammatically correct?

Q2. Is the stem answerable by the text?

Q3. Are all alternatives relevant for the given stem?

Q4. Is the alternative (a) the only correct answer?

If the answer to Q1 is *No*, we report further whether stem, alternatives, or both are ungrammatical. Additionally, we add the category **gibberish** denoting cases when both stem and alternatives are ungrammatical and not formatted properly, or when at least one of

[12] After excluding the special tokens, e.g., `[SEP]`

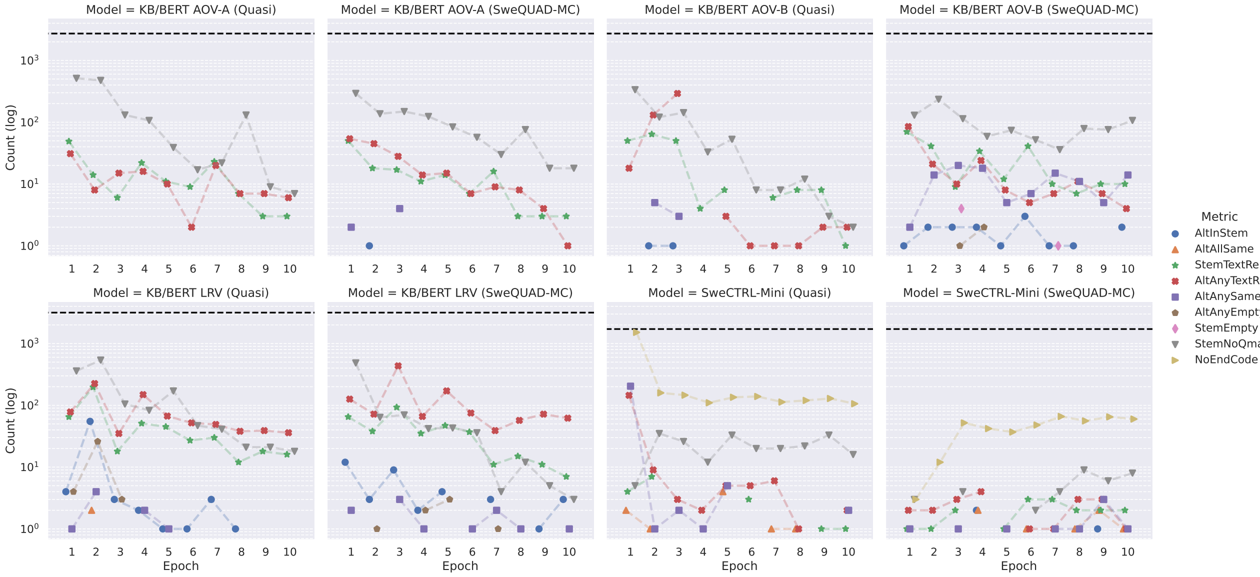


Figure 1: The automatically computed badness metrics for the saved checkpoints of all fine-tuned models. The plots are in *the logarithmic scale* on the y-axis.

	Trained on SweQUAD-MC				**Trained on Quasi**			
	KB/BERT			**SweCTRL**	**KB/BERT**			**SweCTRL**
	LRV	AOV-A	AOV-B		LRV	AOV-A	AOV-B	
Training epochs	8	10	9	5	8	10	6	9

Table 2: The selected checkpoints (denoted by the number of training epochs) based on the automatically calculated badness metrics.

them is not written in valid Swedish (e.g., there are some loose tokens or words that cannot be connected).

If the answer to Q2 is *No*, we investigate the reasons behind the stems being unanswerable. Following Kalpakchi and Boye (2023a), we categorize such stems into **contradictive** (including presuppositions disagreeing with the text), **undiscussed** (inquiring about information not present in the text), or **ambiguous** (not providing enough information to choose one definite alternative).

If the answer to Q3 is *No*, we also investigate the reasons behind it. Following Kalpakchi and Boye (2023a), we categorize the problematic alternatives into **misfocused** or **heterogeneous**. The former category means that at least one of the alternatives does not provide the type of information requested in the stem. For instance, the stem “When was Alfred Nobel born?” with one of the alternatives being “Stockholm” is enough to classify such MCQ as having misfocused alternatives. The latter category means that one or more of the provided alternatives “stick out” and could provide a meta-clue to the students. For instance, the stem “When was Alfred Nobel born?” with the alternatives “21 October 1833”, “1848”, “1792”, “1835” would be classified to have heterogeneous alternatives, since the first alternative (which happens to be the key) is more detailed than the others and thus “sticks out”. Additionally, we introduce two new categories: **empty alternative(s)** (when at least one of the generated alternatives is an empty string), or **duplicate alternatives** (when there are at least two identical alternatives, lowering the effective number of alternatives).

If the answer to Q4 is *No*, we look into three sub-questions to understand why. If any sub-question gets a negative answer, we do not investigate the latter ones. The first sub-question is whether any alternative is the key (the correct answer), to begin with. The second sub-question is whether there is more than one alternative that could be considered to be the key, the case which is referred to as **overlapping alternatives**. The final sub-question is whether the only present key is actually the alternative (a).

Answering Q1 - Q4 required manual annotations, which we did ourselves using an iterative annotation process (annotating – discussing issues – reannotating). We used an instance of Textinator (Kalpakchi and Boye, 2022b) as the annotation tool. The annotation process was blind, meaning that the generated MCQs and their texts were presented in a random order without any indication as to which model they were sampled from (all model-specific tokens and all question numbers were removed). After the evaluation was done, the separately generated key file (previously unseen by the annotator) was used to match the text annotations with their corresponding models.

7.1 Single-sentence texts

Before conducting evaluation on texts from Plugga, taken from the real-world reading comprehension examinations, we turn to a toy domain of extremely small texts consisting of only one sentence. We refer to texts from this toy domain as single-sentence texts (SSTs).

The rationale behind testing the models on SSTs is to facilitate a quick check whether the generated MCQs inquire *only* about the information given in the text. This concern arises from the well-known fact that large language models can generate pieces of text that sound plausible, but are either irrelevant to the given situation or simply false. In our early tests, we noticed that models tend to make up MCQs that are in line with the general topic of the text (e.g. about Sweden) but do not rely on the facts presented in the text. Such artifacts are absolutely unacceptable when producing reading comprehension tasks since the information necessary for answering a stem **must** be present in the text. While the aforementioned checks can be done on any corpus of texts (as we will do in Section 7.2), the idea with SSTs is to make such checks quick and simple. Another advantage with SSTs is that the evaluation becomes more controlled, as we can observe how well models react to slight changes of the text formulations, e.g. whether they are able to pick up slight changes or added information.

For the evaluation in this section we have created the SSTs presented in Table 3, which include 20 *core SSTs* and three *extra SSTs* (marked with asterisks). The extra SSTs contain a specific kind of grammatical errors, namely anglicisms. This is to check whether GPT-based models trained predominantly on English will borrow grammatical constructs from English, even when evaluated on Swedish texts. For every SST, we requested each model to generate **five MCQs** ($N_q^T = 5$) with the first alternative, (a), being correct.

7.1.1 Overview of the results

Our main finding was that two models produced substantially more MCQs without any problems at all, namely ChatGPT (50.43% problem-free MCQs), and GPT-3 (48.7% problem-free MCQs).

An overview of the found problems is presented in Figure 2. The problems in the legend are sorted by the level of their severity, i.e. the harder it is to fix the MCQ, the more severe the problem is. The histogram in Figure 2 accounts only for *the most severe problem* for each MCQ, meaning that the MCQ is guaranteed to not have problems higher in the list, but could still exhibit the problems lower in the list.

To address Q0, the number of results produced by the model, almost all models produced the requested $N_q^T = 5$ per SST. The only exception is the UD baseline that produced substantially fewer MCQs (producing none for the majority of SSTs).

Related to Q0, the models produced different number of alternatives, as reported in Figure 3. Note that vast majority of MCQs contain four alternatives, including both ChatGPT, and GPT-3 for which the number of requested alternatives was unspecified. The only two models with the substantial number of MCQs with other than four alternatives are SweCTRL-Mini trained on SweQUAD-MC (because the training data mostly contained three alternatives), and Zero-shot SweCTRL-Mini.

To address Q1, the issue of grammatical correctness, we look at the three most severe problems from Figure 2. The first and the most severe problem in the list is *gibberish* (red in Figure 2). Gibberish MCQs do not even provide a starting point for fixing an MCQ and require creating a new one altogether, which is why it is the most severe problem. The problem is rare among most of the models, except KB/BERT AOV-B trained on SweQUAD-MC (where it is present for the majority of MCQs).

The next two problems by severity are: *ungrammatical stems* (dark orange in Figure 2), and *ungrammatical alternatives* (light orange in Figure 2). These problems provide a starting point for fixing an MCQ, although still require re-writing major parts of the MCQ. We note that at least one of these two problems is present for every model. The least amount of ungrammatical MCQs were produced by ChatGPT, followed by GPT-3, which is in turn closely followed by KB/BERT AOV-A trained on Quasi.

Strikingly, KB/BERT AOV-B trained on SweQUAD-MC produced *all* MCQs exhibiting one of the three aforementioned problems. For this reason, the model is *excluded* from the further analysis.

Next, we address **Q2, whether or not all stems were answerable by the text**. In fact, the only model with all grammatical stems being answerable by the text is the UD baseline, but it has generated substantially fewer MCQs than the other models. Otherwise, the most frequent reason was that the stems were *undiscussed*, i.e., the answer to the question was not present or inferrable from the text (dark purple in Figure 2). The model with the smallest number of undiscussed stems was GPT-3, followed by ChatGPT. *Contradictive* stems (light purple in Figure 2) and *ambiguous* stems

SST ID	Swedish	English
SST-1	Stockholm är Sveriges huvudstad.	Stockholm is Sweden's capital.
SST-2	Kyiv är Ukrainas huvudstad.	Kyiv is Ukraine's capital.
SST-3	Skranos är Alpongwas huvudstad.	Skranos is Alpongwa's/Alpongwas' capital.
SST-4	Stockholm är Sveriges huvudstad och den största staden i landet.	Stockholm is Sweden's capital and the largest city in the country.
SST-5	Stockholm är huvudstaden och den största staden i Sverige.	Stockholm is the capital and the largest city in Sweden.
SST-6	Kyiv är Ukrainas huvudstad och den största staden i landet.	Kyiv is Ukraine's capital and the largest city in the country.
SST-7	Kyiv är huvudstaden och den största staden i Ukraina.	Kyiv is the capital and the largest city in Ukraine.
SST-8	Skranos är Alpongwas huvudstad och den största staden i landet.	Skranos is Alpongwa's/Alpongwas' capital and the largest city in the country.
SST-9	Skranos är huvudstaden och den största staden i Alpongwa.	Skranos is the capital and the largest city in Alpongwa.
SST-10	Engelska är svårt.	English is difficult.
SST-11	Bengt tycker att engelska är svårt.	Bengt thinks that English is difficult.
SST-12	Anna berättar att Bengt tycker att engelska är svårt.	Anna tells that Bengt thinks English is difficult.
SST-13	Anna är 20 år, Bengt är dubbelt så gammal.	Anna is 20 years old, Bengt is twice as old.
SST-14	Anna är 20 år, Bengt är dubbelt så ung.	Anna is 20 years old, Bengt is twice as young.
SST-15	Bengt är 20 år, Anna är dubbelt så gammal.	Bengt is 20 years old, Anna is twice as old.
SST-16	Bengt är 20 år, Anna är dubbelt så ung.	Bengt is 20 years old, Anna is twice as young.
SST-17 – SST-20 are the same as SST-13 – SST-16, but with the number 20 replaced by 38.		
SST-1*	Stockholm är huvudstaden av Sverige.*	Stockholm is the capital of Sweden.
SST-2*	Kyiv är huvudstaden av Ukraina.*	Kyiv is the capital of Ukraine
SST-3*	Skranos är huvudstaden av Alpongwa.*	Skranos is the capital of Alpongwa.

Table 3: The single-sentence texts (SSTs) used for human evaluation, along with their English translations. Extra SSTs are denoted by asterisks (*).

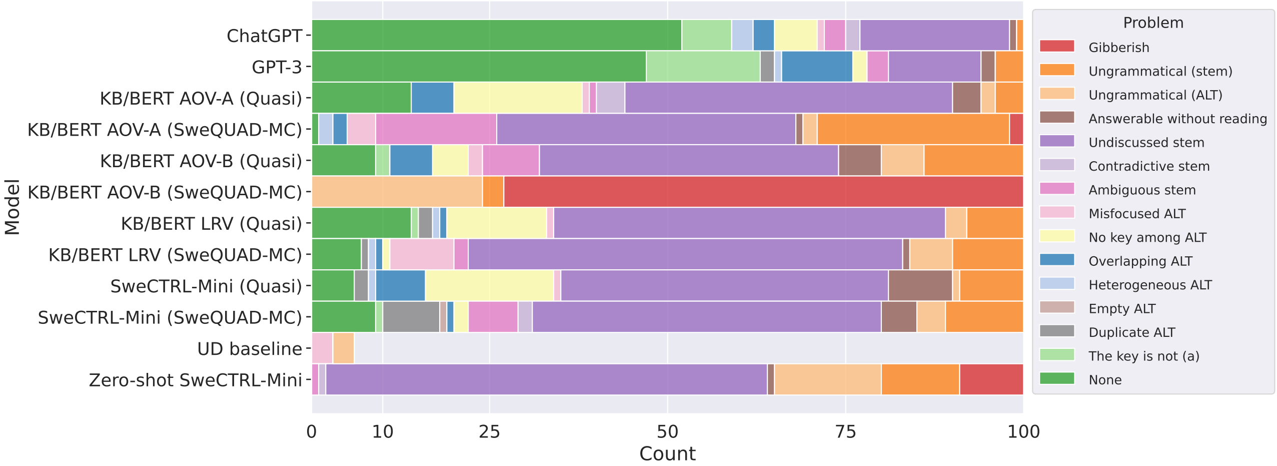


Figure 2: The distribution of problems in the MCQs generated by the 12 evaluated models on the *core* SSTs. The problems are sorted by the severity from the most to the least severe, with *None* (in dark green) indicating the number of MCQs without any aforementioned problems. ALT stands for "alternative(s)".

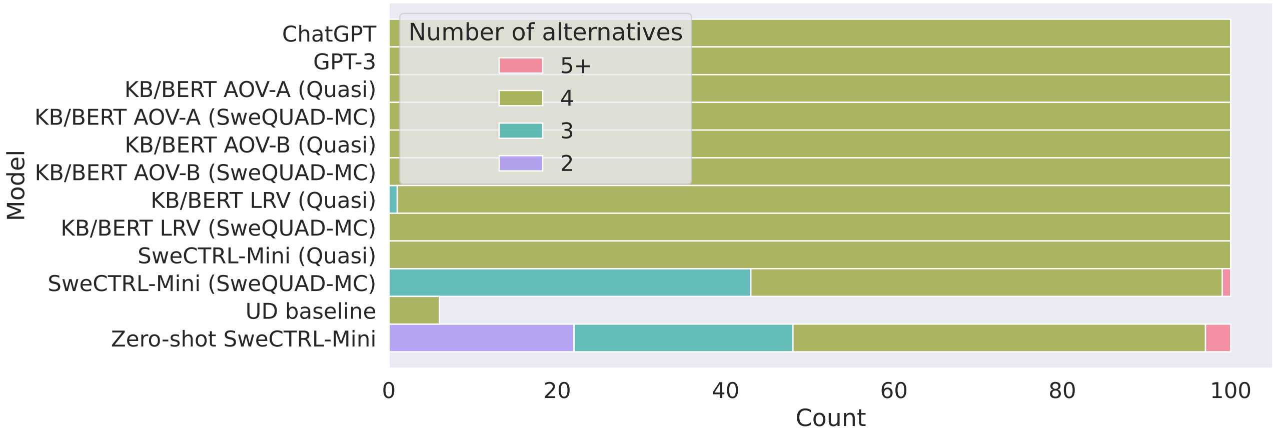


Figure 3: The number of alternatives in the MCQs generated by the 12 evaluated models on the *core* SSTs.

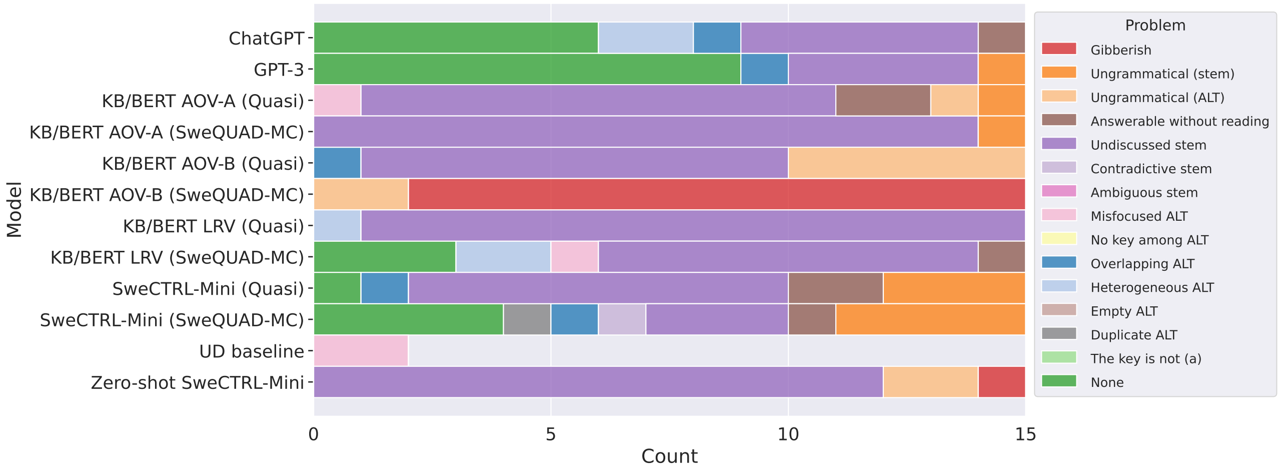


Figure 4: The distribution of problems in the MCQs generated by the 12 evaluated models on the *extra* SSTs (SST-1*, SST-2*, and SST-3*). Gibberish, ungrammatical stems and alternatives do not account for the anglicisms introduced on purpose. ALT stands for “alternative(s)”

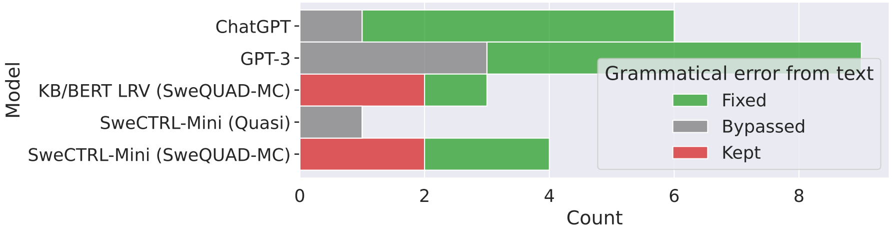


Figure 5: The distribution of *acceptable* MCQs generated by the 12 evaluated models on the *extra* SSTs (SST-1*, SST-2*, and SST-3*) based on whether the introduced anglicisms were fixed, kept or bypassed (by using other formulations).

(dark pink in Figure 2) were much less frequent in comparison (for all models).

Related to Q2, all models produced some MCQs that were answerable without reading the text (dark brown in Figure 2) with SweCTRL-Mini trained on Quasi producing the most such MCQs (by a substantial margin).

To address Q3, whether or not all answer alternatives were relevant for each stem, we note that such problems were infrequent compared to the stem-related problems discussed above. The model with the most MCQs with duplicate alternatives (by a substantial margin) is SweCTRL-Mini trained on SweQUAD-MC, which is also the only model that generated empty alternatives (although for negligibly few MCQs).

Q4 concerns the number of correct answers, of which there should be exactly one, and preferably, this should be alternative (a). This was not always the case: All models except the baselines produced some MCQs with no correct answer at all among the alternatives (light yellow in Figure 2), with the only exception being KB/BERT AOV-A trained on SweQUAD-MC (which had more severe problems for the majority of its MCQs). The two models that produced the most MCQs without a correct answer (in roughly equal amounts) were trained on Quasi, namely KB/BERT AOV-A, and SweCTRL-Mini.

All the models except the baselines also produced MCQs with more than one correct alternative (dark blue in Figure 2). The model with the most such MCQs is GPT-3, whereas the runner-up (with substantially fewer MCQs) is SweCTRL-Mini trained on Quasi.

Finally, as the third and final sub-question of Q4, we checked whether or not the MCQs with exactly one correct alternative indeed had alternative (a) as the key. This was the case for most of the models. Two notable exceptions are GPT-3, and ChatGPT, producing substantially more MCQs with (a) not being the correct alternative (light green in Figure 2).

7.1.2 Error analysis

The distribution of non-problematic MCQs across the single-sentence texts (SSTs) is reported in Figure 6 (only for models that produced at least one such MCQ). Two best-performing models, ChatGPT and GPT-3, have generated at least one acceptable MCQ for almost every SST (except SST-11 for GPT-3). In contrast, generated MCQs for all the other models are distributed more sparsely among the SSTs. The non-GPT model with the best coverage across the SSTs is KB/BERT LRV trained on Quasi with 9 out of 23 SSTs receiving at least one generated MCQ. The very same model is also the model that generated the most acceptable MCQs (14) among non-GPT models. The other model with 14 MCQs is KB/BERT AOV-A trained on Quasi, but it has much worse coverage of only 3 SSTs.

The only two models that generated fully identical MCQs were KB/BERT AOV-A, and KB/BERT LRV, both trained on Quasi. Strikingly, neither ChatGPT, nor GPT-3 produced any fully identical MCQs, despite the fact that the prompt did not require the generated questions to be unique.

Looking closer at the SSTs themselves, we note that **SST-1 to SST-3** follow the same structure "X is Y's capital". The number of acceptable MCQs is the same across these three SSTs for GPT-3, while differs for all the other models. One curious case is the following MCQ generated by ChatGPT based on the SST-1:

> *Vilken stad i Sverige är känt som "Venedig i Norden"?* (*Which city in Sweden is known as "Venice of the North"?*)
> a) *Stockholm*
> b) *Malmö*
> c) *Göteborg*
> d) *Sundsvall*

This is a fully valid MCQ, with (a) being the correct answer, but it has absolutely nothing to do with the actual text of SST-1, which is why it was categorized as having undiscussed stem. This example shows that stems can be undiscussed in many different ways. Sometimes they can relate to the broader topic of the text (e.g., about Stockholm/Sweden) and be entirely valid MCQs in isolation, such as the example above. On the other hand, sometimes they can be completely off topic, such as the MCQ below generated also based on SST-1, but by KB/BERT AOV-B fine-tuned on Quasi.

> *Hur många tandläkare finns det här?*
> (*How many dentists are there here?*)
> a) 2
> b) 3
> c) 5
> d) 10

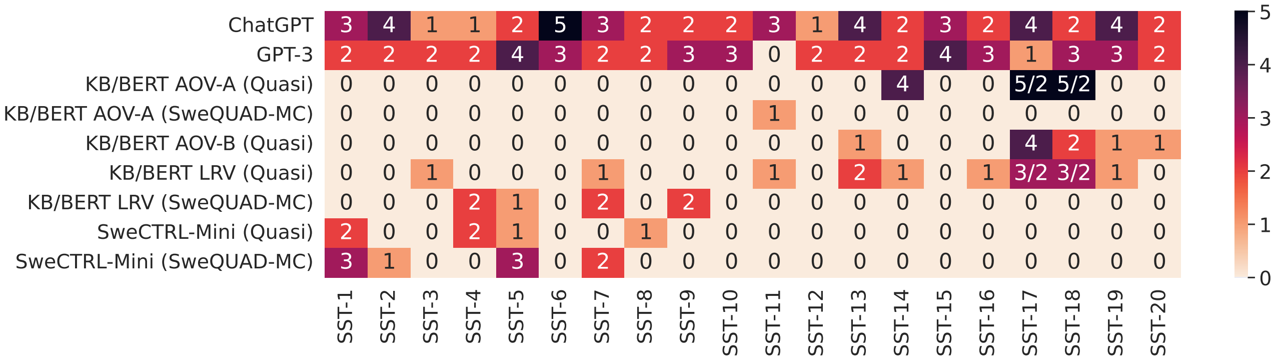
SST-3 involves using made-up toponyms, namely the country Alpongwa, and its capital Skranos. While both ChatGPT, and GPT-3 managed to produce acceptable MCQs, most other models did not. Interestingly, both ChatGPT, and GPT-3 produced acceptable MCQs that used made-up toponyms that sound plausible in their alternatives, as in the MCQ below produced by GPT-3.

> *Vad är huvudstaden i Alpongwa?*
> (*What is the capital in Alpongwa?*)
> a) *Skranos*
> b) *Pangea*
> c) *Malvin*
> d) *Rislanda*

By contrast, the only acceptable MCQ based on SST-3 produced by a non-GPT model, namely KB/BERT LRV trained on Quasi, did not use made-up toponyms:

> *Var ligger Alpongwas huvudstad?*
> (*Where is the capital of Alpongwa?*)
> a) *Skranos*
> b) *Oslo*
> c) *Göteborg*
> d) *Stockholm*

The next batch, **SST-4 to SST-9**, extend SST-1 through SST-3 with one more piece of information. The new structural templates are "X is the capital and the



	SST-1	SST-2	SST-3	SST-4	SST-5	SST-6	SST-7	SST-8	SST-9	SST-10	SST-11	SST-12	SST-13	SST-14	SST-15	SST-16	SST-17	SST-18	SST-19	SST-20
ChatGPT	3	4	1	1	2	5	3	2	2	2	3	1	4	2	3	2	4	2	4	2
GPT-3	2	2	2	2	4	3	2	2	3	3	0	2	2	2	4	3	1	3	3	2
KB/BERT AOV-A (Quasi)	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	5/2	5/2	0	0
KB/BERT AOV-A (SweQUAD-MC)	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
KB/BERT AOV-B (Quasi)	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4	2	1	1
KB/BERT LRV (Quasi)	0	0	1	0	0	0	1	0	0	0	1	0	2	1	0	1	3/2	3/2	1	0
KB/BERT LRV (SweQUAD-MC)	0	0	0	2	1	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0
SweCTRL-Mini (Quasi)	2	0	0	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SweCTRL-Mini (SweQUAD-MC)	3	1	0	0	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 6: The distribution of the generated MCQs with no problems (dark green in Figure 2). The cells with slashes ("/") indicate cases with fully identical MCQs, the format reads "total MCQs with no problems / of which unique".

largest city in Y" for the evenly numbered SSTs, and "X is Y's capital and the largest city in the country" for oddly numbered SSTs. For these examples we note that the number of acceptable MCQs differs between the pairs of reformulations (i.e. SST-4 and SST-5, or SST-6 and SST-7, or SST-8 and SST-9) for all models. Similarly, GPT-based models managed to produce more MCQs overall, although some fine-tuned models perform on-par on these SSTs, except SST-6 and SST-8.

The following MCQ produced by ChatGPT based on SST-4 is an interesting example of an MCQ with heterogeneous alternatives:

> *Vilken stad är större än Stockholm i Sverige?*
> (*Which city is larger than Stockholm in Sweden?*)
> a. *Ingen, Stockholm är den största staden.*
> (*None, Stockholm is the largest city.*)
> b. *Göteborg.*
> c. *Malmö.*
> d. *Uppsala.*

While the alternative (a) is the key, it is clearly longer than all the others. If formulated simply *Ingen* (*None*), then the problem would have disappeared. However, similarly, to undiscussed stems, there are many ways in which the alternatives can be heterogeneous. Additionally, there are different number of alternatives that can "stick out". For instance, in the following MCQ based on SST-9 produced by SweCTRL-Mini trained on Quasi, two alternatives are heterogeneous:

> *Var ligger staden Skranos?*
> (*Where is the city of Skranos?*)
> a) *I Alpongwa* (*In Alpongwa*)
> b) *I huvudstaden* (*In the capital*)
> c) *I Skranos* (*In Skranos*)
> d) *I den kinesiska huvudstaden* (*In the Chinese capital*)

Note that this question is heterogeneous because alternatives (c) and (d) do not use proper names, and the alternative (d) is longer than all the others. Nevertheless, this MCQ was marked as answerable without reading, because the alternative (c) is a correct common-sense alternative simply after reading the stem. Note, however, that the alternative (c) is unlikely to be used in the real reading comprehension tests.

The very same model, SweCTRL-Mini trained on Quasi, produced one of the few acceptable MCQs based on SST-8, which did not use proper names as alternatives.

> *Vilken stad är Alpongwas huvudstad?*
> (*Which city is Alpongwa's capital?*)
> a) *Den största staden* (*The largest city*)
> b) *Den minsta staden* (*The smallest city*)
> c) *Den största floden* (*The largest river*)
> d) *Den högsta punkten* (*The highest point*)

While both (c) and (d) do have nothing to do with cities (and could be viewed as misfocused), such MCQ might still be useful for people just starting to learn the language (which is why it is viewed as acceptable in this evaluation).

Another acceptable MCQ also based on SST-8 without proper names in alternatives was produced by ChatGPT:

> *Hur stor är Skranos jämfört med andra städer i Alpongwa?*
> (*How large is Skranos compared to other cities in Alpongwa?*)
> a. *Störst* (*The largest*)
> b. *Minst* (*The smallest*)
> c. *Andra störst* (*The second largest*)
> d. *Fjärde störst* (*The fourth largest*)

Both this and previous MCQs took advantage of the second clause added to SST-8 compared to SST-3.

Another interesting aspect concerns SST-2, SST-6, and SST-7, namely that the capital of Ukraine has two alternative spellings in English: *Kyiv*, and *Kiev*. All SSTs

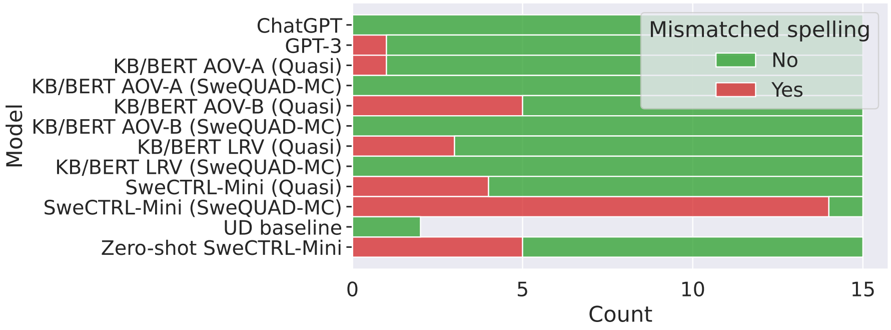


Figure 7: The distribution of the generated MCQs by the presence of mismatched spelling in SST-2, SST-2*, SST-6 or SST-7. The mismatch counts if the spelling of the capital of Ukraine used in the MCQ does not match that of the SST, i.e. Kiev instead of Kyiv (as used in the SSTs).

used the former spelling, but it is relevant to check how well the models comply with the spelling of the text. For that we have manually marked all MCQs that used spelling *Kiev* instead of *Kyiv* present in the SSTs. The distribution of such cases of **mismatched spelling** across the models for the four aforementioned SSTs is shown in Figure 7. We note that the models mostly complied with the spelling of the text, except SweCTRL-Mini trained on SweQUAD-MC which used the alternative spelling most of the time. Interestingly, one of the MCQs based on SST-6 produced by ChatGPT, explicitly asked about the spelling difference:

> *Vad är en annan beteckning för Kyiv?*
> (*What is another term for Kyiv?*)
> a) Kiev
> b) Krakow
> c) Vilnius
> d) Belgrad

Because the nature of the MCQ required the use of both spellings, this was not counted as a mismatch in spellings.

The next batch, namely **SST-10 to SST-12**, are opinion-based statements. Their aim is to test whether the models will add constructs like "according to X", or "as X thinks". We note that most of the models failed to produce an acceptable MCQ on any of these SSTs. Two notable exceptions are ChatGPT and GPT-3 (although GPT-3 also failed on SST-11). Single acceptable MCQs were also produced by two models based on KB/BERT, and one model based on SweCTRL-Mini. In this batch, SST-10 is the interesting one, because the statement sounds like a fact, but it clearly is an opinion, although without specifying whose. The best and only possible kind of MCQ for this SST was produced by ChatGPT:

> *Vilket påstående beskriver författarens åsikt om engelska?*
> (*Which statement describes the author's opinion about English?*)
> a) *Engelska är svårt.* (*English is hard.*)
> b) *Engelska är lätt.* (*English is easy.*)
> c) *Engelska är roligt.* (*English is fun.*)
> d) *Engelska är tråkigt.* (*English is boring.*)

A similar approach was taken by GPT-3 in the following MCQ:

> *Vilken beskrivning är riktig för texten?*
> (*Which description is correct for the text?*)
> a) *Texten beskriver att engelska är svårt*
> (*The text describes that English is hard*)
> b) *Texten beskriver att franska är svårt*
> (*The text describes that French is hard*)
> c) *Texten beskriver att spanska är svårt*
> (*The text describes that Spanish is hard*)
> d) *Texten beskriver att tyska är svårt*
> (*The text describes that German is hard*)

The acceptable MCQs generated on SST-11 and SST-12 took into account that the statements were opinions. For instance, as in the following (and only) acceptable MCQs generated based on SST-11 by KB/BERT LRV trained on Quasi:

> *Vad är det för språk som Bengt tycker är svårt?*
> (*What language does Bengt find difficult?*)
> a) Engelska (*English*)
> b) Svenska (*Swedish*)
> c) Italienska (*Italian*)
> d) Spanska (*Spanish*)

The next batch of SSTs, namely **SST-13 to SST-20**, have to do with counting (understanding who is twice as old, or twice as young). One of the most common mistakes in MCQs for these SSTs had to do with wrong counting. Notably, none of the models based on SweCTRL-Mini were able to produce any acceptable MCQ for this batch. This is also the only batch with absolutely identical acceptable MCQs produced by KB/BERT AOV-A, and KB/BERT LRV, both trained on Quasi.

Additionally, this was the only batch including MCQs with hypothetical stems (all of which were produced by ChatGPT), for instance, the following MCQ based on SST-16:

> *Om Bengt fyller 25 år, hur gammal kommer Anna att vara då?*
> (*If Bengt turns 25, how old will Anna be then?*)
> a) *30 år* (*30 y.o.*)
> b) *15 år* (*15 y.o.*)
> c) *10 år* (*10 y.o.*)
> d) *40 år* (*40 y.o.*)

SST-16 tells us that Bengt is 20, and Anna is twice as young, meaning 10. Then if Bengt turns 25, Anna will also become five years older, meaning 15, hence the correct alternative is (b), and the MCQ was marked as not having (a) as the key. In fact, out of nine generated hypothetical MCQs, only one was acceptable, while three MCQs had ambiguous stems, the other three MCQs did not have (a) as the correct alternative, and the final two MCQs did not provide the key at all. Nevertheless, it is interesting to observe that current state-of-the-art models are capable of producing more challenging MCQs with hypothetical stems. However, one could argue that such MCQs test skills in mathematics, rather than in reading comprehension, a discussion that we will not develop further in this article.

7.2 Plugga

For texts in Plugga we requested N_q^T MCQs per text as calculated by the Equation 1. Both GPT-3 and ChatGPT were able to accommodate the whole input text at once. However, some texts were too long for the context windows of KB/BERT and SweCTRL-Mini (recall that the input for these approaches has to include the input text, *and* as many masked tokens as the final output will contain). In these cases we split the text into multiple parts of L_T tokens each. Often this would result into the last chunk of the text to be be left as a remainder with $< L_T$ tokens. This last chunk could even constitute one sentence, which might not always be enough to generate an MCQ. Hence we took the last L_T tokens as the last chunk meaning that the last and the penultimate pieces of text will overlap. To exemplify, if $L_T = 3$ and the text is `A B C D E F G`, we would split the text into the following chunks: (`A B C`), (`D E F`), (`E F G`). This, together with the fact that the Equation 1 includes rounding up, means that the smaller the L_T, the more MCQs the model will generate compared to the models with larger L_T on the same texts.

7.2.1 Overview of the results

In summary, the GPT-based models, ChatGPT and GPT-3, produced a substantially larger number of acceptable MCQs (63.7% and 37.1%, respectively), compared to the other models. This result is similar to the results of the SST-based evaluation.

An overview of the problems found in the output MCQs is presented in Figure 8, and, similarly to the evaluation on SST, the histogram accounts only for *the most severe problem* for each MCQ. For the sake of brevity, we will omit to list the datasets that the models were trained on in the remainder of the section, since both models trained on KB/BERT were trained on the Quasi dataset, and there is only one version of SweCTRL-Mini.

Similarly to the SST evaluation, we are interested in the very same Q0 - Q4 aspects. Recall that these aspects have been defined as follows:

Q0. Did the model produce the requested number N_q^T of MCQs?

Q1. Are the question (stem) and all alternatives grammatically correct?

Q2. Is the stem answerable by the text?

Q3. Are all alternatives relevant for the given stem?

Q4. Is the alternative (a) the only correct answer?

To address Q0, the number of results produced by the model, only the fine-tuned models generated the requested number of MCQs (shown by the black dashed lines in Figure 8). ChatGPT was slightly short of the target (generating 91.8% of the requested MCQs), while GPT-3 was substantially off the target (generating 33.8% of the requested amount). Concerning the distribution of under-generated MCQs in Figure 10, we note the ChatGPT generated fewer MCQs only for one text due to it reaching the maximum context window size. In contrast, GPT-3 generated fewer MCQs on most of the texts without reaching the maximum context window size for any of the texts.

Related to Q0, the models produced different number of alternatives, as reported in Figure 9. Similarly to SST-based evaluation, vast majority of MCQs contain four alternatives except SweCTRL-Mini that mostly featured MCQs with three alternatives.

To address Q1, the issue of grammatical correctness, all models produced some MCQs with ungrammatical stems and/or alternatives. However, both ChatGPT and GPT-3 produced a very small number of such MCQs. All fine-tuned models generated a larger amount of ungrammatical MCQs compared to GPT-based models, with SweCTRL-Mini even producing a couple of gibberish MCQs. It should be noted that gibberish here included some loose tokens for one of the

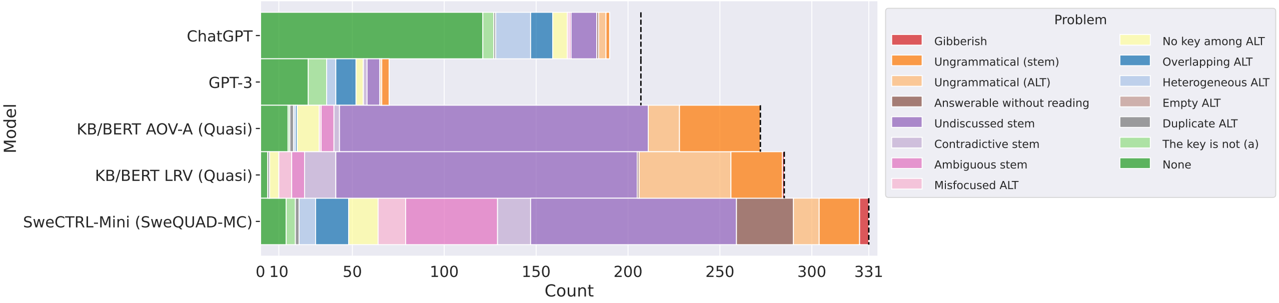


Figure 8: The distribution of problems in the MCQs generated by the TOP-5 best-performing models on SSTs on the texts from Plugga. Black dashed lines indicate the requested number of MCQs to be generated. The problems are sorted by the severity from the most to the least severe, with *None* (in dark green) indicating the number of MCQs without any aforementioned problems. ALT stands for "alternative(s)".

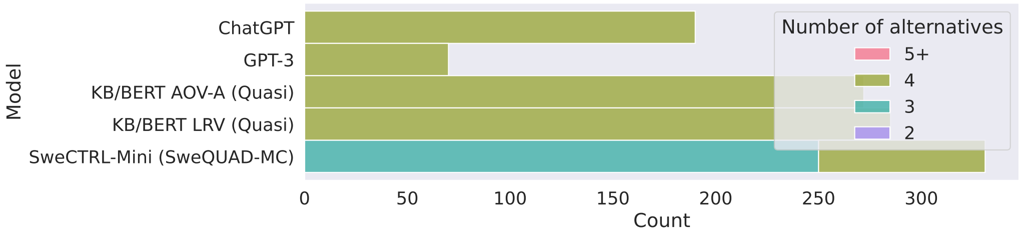


Figure 9: The number of alternatives in the MCQs generated by the 12 evaluated models on the texts from Plugga.

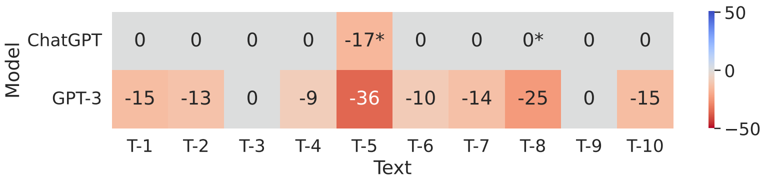


Figure 10: The difference between the actual number of generated MCQs and the requested ones (negative values mean under-generation, positive ones – over-generation, zeros – exactly on point) The cells with the asterisk (*) indicate the cases when the model stopped generation because it reached the maximum context window size.

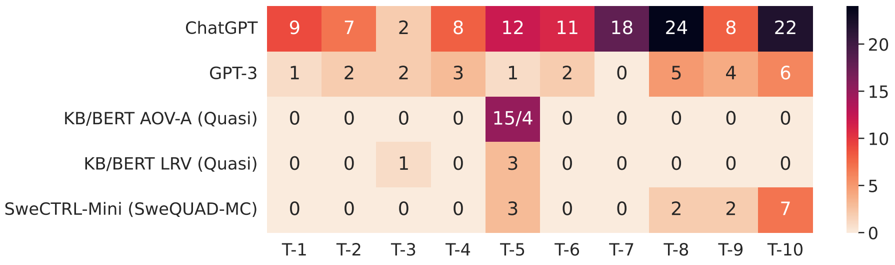


Figure 11: The distribution of the generated MCQs with no problems (dark green in Figure 2). The cells with slashes ("/") indicate cases with fully identical MCQs, the format reads "total MCQs with no problems / of which unique".

alternatives (while the rest of the MCQ is OK), which is radically different from most gibberish encountered in the evaluation on the SSTs previously.

Next, we address **Q2, whether or not all stems were answerable by the text**. In fact, not all stems were answerable by the text, with the most frequent reason (similar to SSTs) being that the stems were undiscussed (dark purple in Figure 8). The pattern is similar to the evaluation on SSTs with the least undiscussed stems being generated by GPT-3, fol-

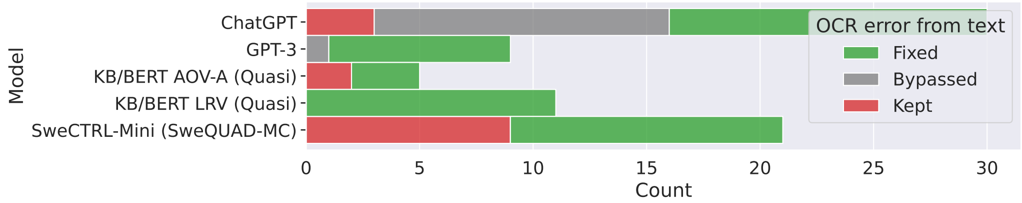


Figure 12: The distribution of *all* MCQs generated by the evaluated models on the Text 8 from Pluga based on whether the OCR error was fixed, kept or bypassed (by using other formulations).

lowed by ChatGPT. Interestingly SweCTRL-Mini generated fewer undiscussed MCQs than models based on KB/BERT, despite producing more MCQs in total. Both contradictive (light purple in Figure 8) and ambiguous (dark pink in Figure 8) stems were much more infrequent in comparison. The largest number of contradictive stems was generated by SweCTRL-Mini (18 MCQs), closely followed by KB/BERT LRV (17 MCQs). The model with a substantially larger amount of ambiguous stems compared to the other models is SweCTRL-Mini.

Related to Q2, one model, namely SweCTRL-Mini, produced substantially more MCQs that are answerable without reading the text (dark brown in Figure 8) compared to the other models. The other two models producing a single such MCQ each were KB/BERT LRV and ChatGPT.

To address Q3, whether all answer alternatives were relevant for each stem, we note that this was not always the case. Similarly to the SST-based evaluation, both misfocused and heterogeneous alternatives are infrequent compared to stem-related problems. The model with the largest number of misfocused alternatives (light pink in Figure 8) is SweCTRL-Mini. The model with the most heterogeneous alternatives (light blue in Figure 8) is still ChatGPT. Additionally, no model produced any empty alternatives, whereas all models except GPT-3 produced at most two MCQs with duplicate alternatives.

Q4 concerns the number of correct answers, of which there should be exactly one, and preferably, this should be alternative (a). This was not always the case: All models produced some MCQs with no correct answer at all among the alternatives (light yellow in Figure 8). The two models that produced the most such MCQs (with roughly equal amounts) were KB/BERT AOV-A and SweCTRL-Mini.

All the models, except KB/BERT LRV, produced MCQs with more than one correct alternative (dark blue in Figure 8).The model with the most such MCQs is SweCTRL-Mini, whereas the runner-ups (with a roughly equal number of such MCQs) are the GPT-based models.

Finally, as the third sub-question of Q4, all models, except KB/BERT LRV, produced some MCQs with the correct alternative not being (a) (and without any more severe problems).

7.2.2 Error analysis

The distribution of non-problematic MCQs across the texts in Plugga is presented in Figure 11. The two best-performing models, ChatGPT and GPT-3, have generated at least one acceptable MCQ for almost every text (except Text 7 for GPT-3). Similarly to the SST-based evaluation, the acceptable MCQs generated by all the other models are distributed more sparsely among the texts. The non-GPT model with the best coverage across the texts in Plugga is SweCTRL-Mini with 4 out of 10 texts resulting in at least one generated MCQ. The only model that generated several MCQs that were completely identical is KB/BERT AOV-A, as it also did for the SST-based evaluation. With this in mind, SweCTRL-Mini is also the best non-GPT model when it comes to the number of unique generated MCQs.

Most of the generated MCQs asked about facts or details presented in the text, with only two MCQs, both produced by ChatGPT, requiring high-level text-based inference. One such MCQ based on the Text 8 is presented below:

> *Vad är budskapet i denna historia?*
> (*What is the message in this story?*)
> a) *Det är viktigt att hjälpa andra* (*It is important to help others*)
> b) *Det är bäst att inte bry sig om andra* (*It is best to not care about others*)
> c) *Det är farligt att hjälpa främlingar* (*It is dangerous to help strangers*)
> d) *Det är bäst att vara egoistisk* (*It is best to be selfish*)

Finally, text 8 from Plugga (a news article about the boy, *Josef*, who helped to save the girl in a wheelchair from a snow trap), we purposefully kept one small OCR error that misspelled the boy's name to *Joset**. In contrast to the spelling example from the SST-based evaluation (*Kyiv vs Kiev*), here using the spelling from the text is undesirable. Only three evaluated models managed to generate at least one MCQ for this text, and the distribution of all such MCQs (not only the acceptable ones) based on whether the OCR error was fixed is reported in Figure 12. Notably, the majority of MCQs either fixed or bypassed the misspelling problem by using different formulations. Three models that kept the erroneous spelling are SweCTRL-Mini (for around half of the *its* MCQs), ChatGPT, and KB/BERT AOV-A.

8 Discussion

We have observed a substantial performance difference between ChatGPT (gpt-3.5-turbo-0301) and all other models for the MCQ generation task. We conducted only very limited manual hyper-parameter tuning for the non-GPT models when fine-tuning, which means that there is a possibility that the performance of these models could be boosted further. However, as the datasets used for fine-tuning are quite small, it is unlikely that such a boost will be enough to reach the performance of ChatGPT, which produces acceptable MCQs at a rate at least three times higher.

Additionally, one should bear in mind that the focus of this evaluation was mainly on uncovering **foundational problems** with the generated MCQs. This means that even MCQs judged to be acceptable might not necessarily be of sufficient quality to be useful in a high-stakes testing scenario (such as national exams). For instance, the following MCQ produced by ChatGPT on text 8 was marked as acceptable.

> *Vilken funktion hade tjejens rullstol?* (*Which function did the girl's wheelchair have?*)
> a) *Den var elektrisk* (*It was electrical*)
> b) *Den var manuell* (*It was manual*)
> c) *Den var gjord av trä* (*It was made of wood*)
> d) *Den hade ingen funktion* (*It didn't have any function*)

While, formally, this MCQ did not have any of the problems brought up in this work, it asks about a very minor detail. The type of the wheelchair is mentioned twice in the text but really has nothing to do with the plot of the story. While this MCQ could be useful to test how attentive the reader was, it is not aimed at assessing reading comprehension skills but rather memory, which might also be desirable depending on the target audience for the test. Furthermore, alternative (c) is arguably not plausible, since being made of wood is not a function. Nevertheless, such an alternative might be deemed relevant depending on the target audience of the test (for instance, at the entrance-level classes).

Similarly, the use of "none of the above" alternatives, such as in the MCQ below generated by ChatGPT on Text 7, might not necessarily be desirable.

> *Vad hoppas Fredrika på angående sitt jobb?* (*What does Fredrika hope for regarding her job?*)
> a) *Att hon får ett jobb som känns meningsfullt, även om det inte är så bra betalt* (*That she will have a job that feels meaningful, even if it is not very well paid*)
> b) *Att hon får ett jobb som ger henne hög lön* (*That she will have a high-paying job*)
> c) *Att hon får ett jobb som hon inte behöver engagera sig i* (*That she will have a job that requires minimal effort*)
> d) *Ingenting nämns om vad hon hoppas på angående sitt jobb* (*Nothing is mentioned about what she hopes for regarding her job*)

In fact, Haladyna et al. (2002) report that opinions are split about using alternatives such as (d), which is why such MCQs might not necessarily be judged as acceptable in the larger-scale evaluation.

9 Conclusion

In this article, we have compared the MCQ-generating capabilities of 12 different models, eight models fine-tuned by ourselves, two baselines, as well as GPT-3 (*text-davinci-003*), and ChatGPT (*gpt-3.5-turbo-0301*).

The GPT-based models perform better than the rest of the models. ChatGPT performs substantially better than *all* other models in *all* evaluation settings. GPT-3 performs substantially better than non-GPT models when tested on single-sentence texts, whereas the performance gap on the texts from SFI national tests (the Plugga corpus) is less pronounced in comparison to other models (excluding ChatGPT).

Additionally, we noticed that GPT-based models have an inductive bias for producing MCQs with four alternatives, even when the number of alternatives is unspecified in the prompt.

In our limited experiments with the introduced grammatical errors, we noticed that GPT-based models avoid using anglicisms even if they were introduced in the original text, while other models tend to stick to the text much more frequently. The behavior of the models was less conclusive for the introduced OCR error, except for GPT-3 (which produced substantially fewer MCQs than requested though) and KB/BERT LRV trained on Quasi that did not use the formulation with the OCR error.

At the same time, when sticking to the text was necessary, such as in the example with the alternative spellings of the capital of Ukraine, most models used the spelling from the text with ChatGPT, and some models based on KB/BERT taking the lead. All models based on SweCTRL-Mini were lagging behind and used the formulation that was more frequent in their training data (148 thousand occurrences of the word *Kiev* vs 2052 occurrences of the word *Kyiv*)[13].

In summary, all conducted evaluations point to the fact that the best model for generating MCQs in Swedish is ChatGPT, followed by GPT-3, followed by SweCTRL-Mini fine-tuned on the SweQUAD-MC corpus. The results based on the toy domain of single-sentence texts (SSTs) closely resemble those on the larger-scale texts from the SFI national exam. This indicates that SST-based evaluation might be a viable lower-cost alternative to the full-scale human evaluation, and warrants more extensive studies on the strength of the correlation, and the extent of the time savings.

References

Araki, Jun, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1125–1136, Osaka, Japan. The COLING 2016 Organizing Committee.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chung, Ho-Lam, Ying-Hong Chan, and Yao-Chung Fan. 2020. A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4390–4400, Online. Association for Computational Linguistics.

Guo, Qi, Chinmay Kulkarni, Aniket Kittur, Jeffrey P Bigham, and Emma Brunskill. 2016. Questimator: Generating knowledge assessments for arbitrary topics. In *IJCAI-16: Proceedings of the AAAI Twenty-Fifth International Joint Conference on Artificial Intelligence.*

Haladyna, Thomas M. 2004. *Developing and validating multiple-choice test items.* Routledge.

Haladyna, Thomas M, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.

Kalpakchi, Dmytro and Johan Boye. 2021. BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 387–403, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Kalpakchi, Dmytro and Johan Boye. 2022a. Automatically generating question-answer pairs for assessing basic reading comprehension in Swedish. *arXiv preprint arXiv:2211.15568.*

Kalpakchi, Dmytro and Johan Boye. 2022b. Textinator: an internationalized tool for annotation and human evaluation in natural language processing and generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 856–866, Marseille, France. European Language Resources Association.

Kalpakchi, Dmytro and Johan Boye. 2023a. Quasi: a synthetic question-answering dataset in Swedish using GPT-3 and zero-shot learning. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 477–491, Tórshavn, Faroe Islands. University of Tartu Library.

Kalpakchi, Dmytro and Johan Boye. 2023b. Quinductor: A multilingual data-driven method for generating reading-comprehension questions using universal dependencies. *Natural Language Engineering*, page 1–39.

Kalpakchi, Dmytro and Johan Boye. 2023c. SweCTRL-Mini: a data-transparent Transformer-based large language model for controllable text generation in Swedish. *arXiv preprint arXiv:2304.13994.*

Loshchilov, Ilya and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations.*

Majumder, Mukta and Sujan Kumar Saha. 2015. A system for generating multiple choice questions: With a novel approach for sentence selection. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 64–72, Beijing, China. Association for Computational Linguistics.

Malmsten, Martin, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT.

[13] Checked using the publicly available website for SweCTRL-Mini: https://swectrl.dev/data

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

OECD. 2019. *PISA 2018 Assessment and Analytical Framework*. OECD Publishing, Paris.

OECD. 2021. *PISA 2025 Foreign Language Assessment Framework*. OECD Publishing, Paris.

Offerijns, Jeroen, Suzan Verberne, and Tessa Verhoef. 2020. Better distractions: Transformer-based distractor generation and multiple choice question filtering. *arXiv preprint arXiv:2010.09598*.

Qiu, Zhaopeng, Xian Wu, and Wei Fan. 2020. Automatic distractor generation for multiple choice questions in standard tests. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2096–2106, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhang, Ruqing, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2021. A review on question generation from natural language text. *ACM Transactions on Information Systems (TOIS)*, 40(1):1–43.

Zhou, Xiaorui, Senlin Luo, and Yunfang Wu. 2020. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9725–9732.