

The Future of Natural Language Generation in the Age of Large Language Models

Eduardo Calò, Amazon, The Netherlands educal@amazon.nl*

Saad Mahamood, Shopware, Germany saad@saad.me.uk

Letter abstract Natural language generation (NLG) research has focused in the past on mainly two aspects: (i) the process of language generation in itself (e.g., to give theoretical insights into models of human language) and (ii) the use of such approaches to solve applied problems (e.g., generation of medical reports, weather forecasts, etc.). Since the advent of large language models (LLMs), given their “generation capabilities”, there has been a tendency among researchers to use LLMs for an increasing number of novel tasks. The use of LLMs raises questions for NLG researchers about which tasks, both theoretical and practical, are still considered NLG tasks per se, and which are still relevant, given that LLMs have solved many former problems in language generation. Through a structured literature search, we investigate domain shift in NLG before and after LLMs and discuss the positive and negative aspects that arise from this, with the aim of providing researchers with greater clarity on where we stand as a field.

1 Introduction

The field of natural language generation (NLG) has evolved significantly over the past 40 years. However, the aim of enabling computers to take underlying sources of data and generate textual output systematically (Reiter and Dale, 2000; Gatt and Krahmer, 2018; Reiter, 2025a) has remained constant. This focus on how to get machines to generate language optimally for readers has led to many sub-areas of research focus within the field, where aspects of psycholinguistics are also taken into consideration (Guhe, 2007). In addition to how language should be generated, NLG has also found applications in diverse applied settings, such as weather forecasting (Reiter et al., 2005; González Corbelle et al., 2022), tutoring systems (Di Eugenio et al., 2002; Mpagouli and Hatzilygeroudis, 2009; Maurya and Kochmar, 2025), and healthcare reporting (Portet et al., 2009; Gatt et al., 2009; Mahamood and Reiter, 2011; Hunter et al., 2012; Lyu et al., 2026).

Historically, rule-based systems dominated NLG research, relying on the data-to-text architecture (Reiter, 2007a) to separate the process of generating text into a series of discrete steps that transform data into knowledge/facts and then into narrative by selecting the most relevant aspects to summarize. While these systems were robust and interpretable, their domain specificity limited their generalizability (Karanikolas

et al., 2025). This limitation was broadly recognized by the NLG research community and sought to be addressed with efforts such as semi-supervised learning and deep learning (Weston et al., 2008). End-to-end approaches were explored as an alternative to highly domain-specific rule-based NLG systems, initially employing neural LSTM and sequence-to-sequence models (Wen et al., 2015; Dušek and Jurčiček, 2016; Mei et al., 2016; Gehrmann et al., 2018) and later transformer-based models (Liu et al., 2018; Song et al., 2019; Keskar et al., 2019; Hoang et al., 2019).

Recently, large language models (LLMs) have taken the natural language processing (NLP) landscape by storm, leading researchers to explore their capabilities across a wide range of tasks. NLG has been no exception, impacting both academia and industry (Dale, 2023). Since the release of GPT-3 (Brown et al., 2020), the increasing capabilities of LLMs have led to their adoption for language generation, also blurring traditional boundaries, e.g., by reframing tasks that were previously not considered generative problems, such as analogy resolution (Das and Balke, 2025).

Moreover, many problems previously treated as distinct research tasks, such as content determination (Barzilay and Lapata, 2005) and surface realization (Gatt and Reiter, 2009), are now often subsumed within the use of a single generative model. This raises a pressing question in the NLG field: has the widespread adoption of LLMs for aspects once studied as separate re-

*The work was conducted before joining Amazon.

search problems led to a fundamental change in the nature of research conducted by NLG researchers?

The rapid adoption of LLMs has also brought to the forefront a number of challenges regarding evaluation. The topic was actively explored within the NLG community, with findings such as researchers favouring intrinsic over extrinsic metrics (Gkatzia and Mahamood, 2015) and exploring methodological shortcomings in human evaluations (van der Lee et al., 2019; Howcroft et al., 2020; Belz et al., 2020). However, LLMs introduced more prominently issues of output faithfulness and hallucinations (Maynez et al., 2020), as well as the difficulty of evaluating generative systems reliably, both automatically and through human judgments (Chang et al., 2024). Additional concerns have been raised regarding data contamination (Balloccu et al., 2024), privacy, and broader ethical implications (Zubiaga, 2024). At the same time, the reliance on increasingly large, often closed, models has prompted questions about the sustainability of scaling-based progress and about the reproducibility and methodological soundness of research that depends on systems with limited transparency (Rogers, 2023; Friedrich, 2025).

To better understand the impact of the use of LLMs in NLG research, we compare NLG research pre- and post-LLMs by sampling papers from the ACL Anthology database to see if there is a change in the distribution of generation research tasks performed by researchers. Whilst there has been interest in surveying the broad trends of how LLMs have impacted research (Lee et al., 2023; Movva et al., 2024; Kostikova et al., 2026), we focus on determining what the consequent trends are for NLG specifically as a research field. Has the introduction of LLMs led to a spike in interest in certain problems over others? What are the positive or negative implications of these changes? Are there areas of research in NLG that are now left under-researched or considered “solved” by researchers? Our discussion highlights the emergence of some tasks and the relative decline of others, and reflects on the possible motivations behind these shifts. We also note the growing prominence of evaluation and emphasize the importance of addressing broader concerns, such as reproducibility and data-related issues.

2 Literature Search

To explore changes in NLG research, we conducted a keyword-based literature search with the goal of identifying which NLP tasks occur in papers related to the NLG or LLM domains.¹ We deliberately considered the full range of NLP tasks, because assessing whether NLG

¹Keyword-based searches have been used by other practitioners to survey trends within NLP, such as work done by Reiter (2025b) or Karamolegkou et al. (2025).

has expanded beyond its traditional scope requires examining it in relation to the broader NLP landscape.

To compare the difference in task proportions between pre- and post-LLM ages, the search was carried out in two time periods: 1965-2020 and 2021-2025. As the dividing year, we chose 2020, the publication year of GPT-3. Drawing the exact line between pre- and post-LLM ages can be hard. For instance, some might argue that the release of ChatGPT in 2022 would make a better inflection point. However, we chose 2020 because the publication of GPT-3 already marked a major shift in the scale and capabilities of language models. This shift is reflected in the ACL Anthology, where the first clear increase in research papers mentioning GPT appears in 2021 (Gururaja et al., 2023).

2.1 List of NLP Tasks

To ensure sufficient coverage and appropriate granularity of NLP tasks, we gathered multiple sources, and manually selected and aggregated the tasks listed in them. The resources are the following: Reiter and Dale (2000), Hedderich et al. (2021), Jurafsky and Martin (2025), Vatsal et al. (2025), Qin et al. (2025), and the Wikipedia article on Natural Language Processing. We made sure not to remain too broad, for instance, by excluding broad tasks such as “natural language understanding”, and not to be too fine-grained, for instance, by avoiding very specific tasks such as “aspect sentiment triplet extraction”. We also took care to include spelling, synonymous, and acronymic variants.²

2.2 Methodology

To better understand the research trends of NLG across different research tasks, we performed a search of relevant papers. This was done by downloading the ACL Anthology bibliography file on December 10, 2025, which contained a total of 118,471 papers. Because many papers published before 2020 lack an abstract field in this file, we supplemented the data with information from Rohatgi et al. (2023).³ This resource provides metadata for ACL Anthology papers up to 2022, including abstracts that were extracted directly from the PDF files. After this process, abstracts are available for 90.55% of the papers published between 1965 and 2020, and for 91.78% of the papers published between 2021 and 2025, ensuring a comparable coverage.

From the supplemented bibliography file, we parsed each paper reference to retrieve all occurrences of the

²E.g., “content determination” was also matched with “content selection”, “content planning”, and “content specification”. The codebase, including also the complete list of NLP tasks considered, can be found at: <https://github.com/Eduardo-Calo/nlg-future>.

³We used the file <https://huggingface.co/datasets/WINGNUS/ACL-OCL/blob/main/acl-publication-info.74k.v2.parquet>.

keywords (1) “NLG” alone, and (2) “NLG” together with “LLM” (including their spelling variants), where the keywords can appear in the title, abstract, book title, or URL.⁴ This allowed us to identify papers relevant to our two domains: (1) **NLG**, and (2) **NLG+LLM**.⁵ We then computed the *proportion*⁶ of these papers that also mention an NLP task from §2.1 in their title or abstract. This method covers cases where keywords can surface either as a textual keyword or through a venue name, for example, when locating papers on “paraphrasing”, where “NLG” may appear in the paper (title or abstract) or in the conference name (book title or URL).⁷

To assess the reliability of our keyword-based methodology, we manually validated a stratified sample of 60 papers: 15 papers with task matches and 15 without, from each of the periods 1965-2020 and 2021-2025. The two authors independently annotated whether each paper actually addressed the matched task, achieving a Cohen’s κ of 0.927 (58/60 agreements). After excluding the two disagreements, the keyword-based matches achieved an accuracy of 0.862 against the human annotations. All errors were false negatives, where relevant tasks are discussed but not explicitly mentioned in the title or abstract (e.g., [Amato and Arrita \(2023\)](#), who describes tasks such as surface realization and multilingual generation without naming them explicitly).⁸ Overall, the validation indicates that our methodology is sufficiently reliable for our analysis.

3 Results

Table 1 provides an overview by domain and period. The data show a clear surge in NLG publications in the last four years, with the papers published between 2021 and 2025 amounting to roughly two-thirds of the total output accumulated over more than five decades. Within this recent literature, the NLG+LLM domain accounts for 310 papers, corresponding to roughly one quarter of all NLG papers published in the same period. This indicates that LLMs are used as a core technology in approximately one out of four recent NLG studies. The

⁴We acknowledge that papers may be about NLG or LLMs even without the presence of these keywords. However, the choice of these keywords helps to ensure that the selected papers are actually about these topics.

⁵The NLG+LLM domain is a strict subset of the NLG domain.

⁶We did not use absolute counts because the number of papers between domains and time periods varies.

⁷We acknowledge that some tasks such as “machine translation” have often been treated as separate research areas and not necessarily framed as generation tasks. As a result, papers on these tasks may not explicitly mention the “NLG” keyword, which could lead to an underrepresentation of such work in our data.

⁸Although our manual validation did not identify any false positives, they remain a possible limitation of the keyword-based methodology. For example, a paper may mention “machine translation” in the abstract only to provide context, rather than because machine translation is itself the main task.

proportion of papers focusing on the NLP tasks described in §2.1 is broadly comparable between NLG and NLG+LLM domains and across time periods. Moreover, the data show that individual papers frequently address more than one task, as reflected by the average number of task matches per paper reported in the last row.

Several interesting patterns emerge from Figure 1. Some tasks, such as machine translation, data-to-text, and paraphrasing, have remained stable over time, serving as cornerstones of NLG research. In contrast, classical NLG tasks, like surface realization, content determination, and referring expression generation, have sharply declined, showing that certain traditional tasks have become less central in recent research. Text summarization has also seen a substantial increase in attention. More notably, entirely new tasks are emerging that were not previously within the scope of NLG research, including text classification and mathematical reasoning. Evaluation itself is also gaining prominence, seeing marked jumps between the two periods, a trend closely tied to the rise of hallucination concerns.

Breaking down the trends of the ten most frequent tasks into five-year intervals provides a more fine-grained view of these developments (Figure 2). Patterns emerge from approximately 1981, becoming more pronounced after 1986, with sparse earlier periods. For example, the spike in paraphrasing in 1971-1975 is an artifact of data sparsity, as only three papers fall into that interval and one is on paraphrasing. Evaluation shows a steady increase from 1995, with a further acceleration after 2002. Data-to-text gains prominence after 2001. Conversely, traditional NLG subtasks decline after 2015. Finally, research on text summarization and hallucination detection grows after 2016.

4 Discussion

Our numbers show that the NLG field is drastically changing and moving forward following the latest technological advancements, as illustrated for the tasks in Figure 2. Some tasks like data-to-text have seen increases that coincide with developments such as the pipeline architecture for building NLG systems ([Reiter, 2007a](#); [Reiter and Dale, 2000](#)) and the emergence of “big data” in the early 2000s ([Diebold, 2012](#)). On the other hand, we see significant declines in surface realization, which could be attributed to a shift in interest as neural approaches come to the fore. In what follows, we will try to distill the numbers and explore the positive, negative, and broader implications our results suggest.

4.1 Positive Implications

From our results, there are several positive implications from changes observed pre- and post-LLM ages

Statistic	NLG (1965–2020)	NLG (2021–2025)	NLG+LLM (2021–2025)
Total papers	1,671	1,117	310
Papers with at least one task match	839 (50.21%)	685 (61.32%)	201 (64.84%)
Total task matches	1,172	983	267
Average task matches per paper	1.40	1.44	1.33

Table 1: Overview of papers and task match statistics in the NLG and NLG+LLM domains across time periods. Percentages are computed with respect to the total number of papers in each domain.

for NLG research. Firstly, there is a sharp increase in the number of publications, as can be seen in Table 1. One possible explanation for this growth is that specific NLG tasks have gained renewed prominence or emerged anew in the age of LLMs, as reflected in Figure 1. Tasks such as story generation and code generation show substantial increases in attention, while entirely new ones, like text classification and mathematical reasoning, appear as NLG tasks only since 2021. This may be an indication that the concept of NLG might be broadening due to possible convergence in terms of methods with the use of LLMs. Yet, the emergence and growth of these tasks is encouraging because many of them can be considered of practical use, with the potential to support real-world applications in areas such as education (Cambaz and Zhang, 2024), healthcare (Guo et al., 2024), and engineering (Dong et al., 2025).

One reason tasks that were previously considered unrealistic are now achievable is the confluence of increased compute power and larger datasets, combined with advances in language model architectures (Hagos et al., 2024). In particular, generative pre-trained transformers and their successors provide a unified framework that can be fine-tuned or prompted for a wide range of tasks (McCann et al., 2018; Brown et al., 2020; Min et al., 2021; Liu et al., 2022), making it easier not only for researchers but also for broader communities to apply NLG methods. This pattern mirrors similar developments seen previously in the field of computer vision, where the advent of deep convolutional models (Krizhevsky et al., 2012) led to rapid growth in both research activity and practical applications (Se, 2024).

Another area showing substantial growth is evaluation, which has received steadily increasing attention since the mid-1990s (Figure 2), with further acceleration after 2002, possibly coinciding with the introduction of the BLEU metric (Papineni et al., 2002), and another rise after 2020 (Figures 1 & 2). We interpret this second uptick as a consequence of the increased difficulty of assessing the outputs of LLMs. Compared to earlier approaches, the errors produced by these models tend to be more subtle, which has motivated the development of new or more robust evaluation methods to address challenges related to assessing output quality, particularly with respect to faithfulness and factuality (Maynez et al., 2020; Augenstein et al., 2024). This

trend is closely linked to the growth of research on hallucination, as can be seen by the subsequent interest in hallucination detection post-2020 in Figures 1 & 2. The generation of factually incorrect or unsupported content has emerged as a major concern (Ji et al., 2023; Bouyamourn, 2023; Ravichander et al., 2025; Huang et al., 2025), also with focused studies targeting specific areas of NLG, such as data-to-text (Dušek and Kasner, 2020; Thomson and Reiter, 2020; van Deemter, 2024). These patterns have also been highlighted in other recent studies that emphasize the limitations of existing evaluation methods when applied to LLM-generated text (Schmidtova et al., 2025).

4.2 Negative Implications

Since 2021, roughly one quarter of papers have been leveraging LLMs as part of NLG research (Table 1). This indicates that while LLMs have had a big impact, there is still NLG research happening outside the scope of LLMs. Yet, not all tasks in NLG have seen increases post 2021. Some traditional NLG tasks, such as referring expression generation, surface realization, and content determination, have either declined or stopped being addressed altogether (Figures 1 & 2). These declines could be a natural consequence of researchers considering certain problems within NLG as “solved”, with powerful LLMs able to produce very fluent text in a variety of contexts and domains (Chang et al., 2024), hence solving tasks that once needed bespoke approaches.

However, treating these more linguistically-oriented tasks as solved and no longer investing research effort in them carries significant risks. In particular, it may lead to reduced attention to domains that deal with low-resource languages or resources (Howcroft and Gkatzia, 2022). LLMs have been shown to underperform in low-resource settings and scenarios (Dargis et al., 2024; Azime et al., 2025), and shifting focus entirely towards newer tasks may result in longstanding linguistic challenges being overlooked. Continued work on these foundational tasks remains crucial, especially because support for low-resource languages is lacking (Occhini et al., 2026) yet essential for inclusive language technology for large portions of the global population (Gala et al., 2023; Khan et al., 2024; Issaka et al., 2025; Ojo et al., 2025).

The Future of Natural Language Generation in the Age of Large Language Models

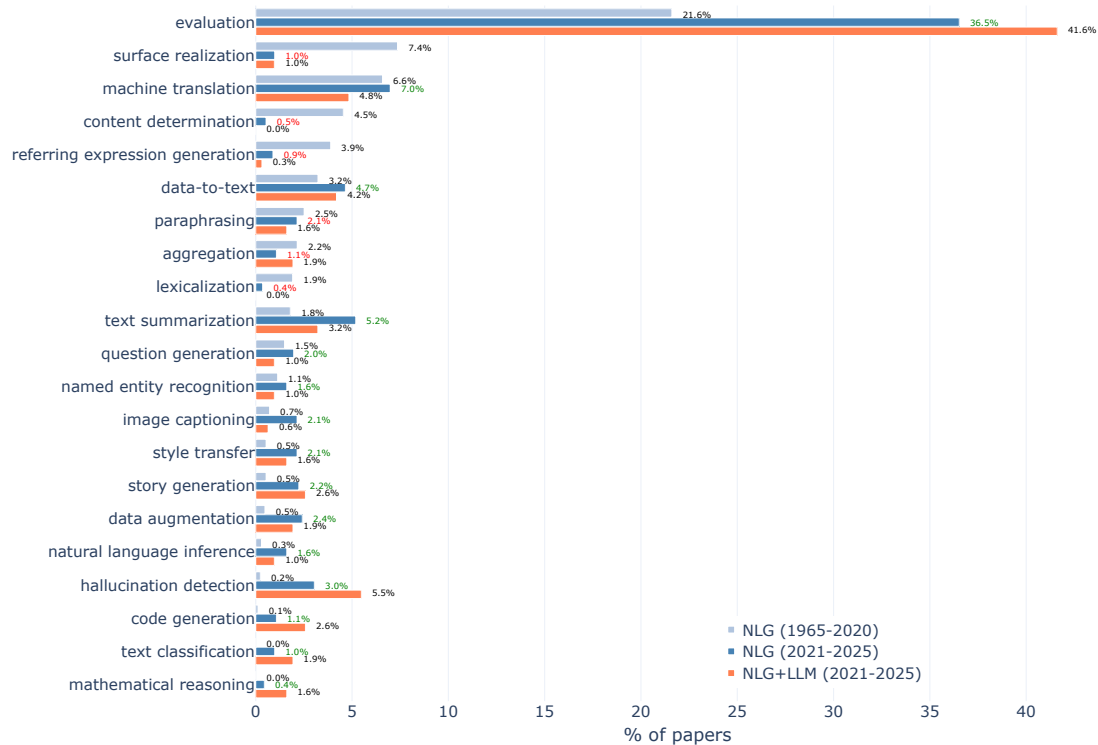


Figure 1: Focus of research in the NLG and NLG+LLM domains across time periods. Percentages indicate the share of papers addressing each task in the respective domain/period. Numbers in green/red indicate tasks that have increased/declined between NLG (1965–2020) and NLG (2021–2025). Only tasks with a representation of at least 1% in at least one bar are shown.

Another concern is whether LLMs are truly the most appropriate technology for every task. For instance, research on referring expression generation has shown that more traditional, non-neural approaches can still remain competitive in certain settings (Same et al., 2022). Similar observations have been made in data-to-text. In the 2023 WebNLG shared-task on low-resource languages (Cripwell et al., 2023), the submission for Irish described by Mille et al. (2023) relies on a rule-based system that achieves competitive performance compared with submissions based on neural approaches. Broader analyses also suggest that although LLMs are highly versatile and perform strongly across many benchmarks, they still fall short of fully solving the breadth of NLP tasks robustly and reliably (Qin et al., 2023; Basmov et al., 2024; Berglund et al., 2024).

Whether this trend is ultimately beneficial for NLG research depends on how one interprets it. On the one hand, declining interest in some classical tasks may indicate that research effort is being redirected toward pressing challenges such as evaluation and faithfulness. On the other hand, it could possibly signal the emergence of a methodological orthodoxy in which LLM-based approaches dominate to such an extent that

alternative, and sometimes more efficient or better-suited, methods are no longer actively explored.

4.3 Broader Implications

The adoption of LLMs within research (Kostikova et al., 2026) has also brought to the forefront broader concerns that remain unresolved. The growing reliance on LLM-based approaches, where models function as black box systems whose internal mechanisms are difficult to inspect (Calderon and Reichart, 2025), raises pressing issues of reproducibility (Branco et al., 2020; Belz et al., 2021, 2022; Belz and Thomson, 2023, 2024; Belz et al., 2025), which lies at the core of reliable evaluation, i.e., the fundamental basis for determining whether an NLG system performs well (Reiter, 2026). These concerns have become even more salient in light of our findings, which show the increasing prominence of evaluation-related research (Figures 1 & 2).

Other concerns relate to the data used to train LLMs. Data contamination can undermine both fair and appropriate benchmarking, calling into question the validity of reported performance gains (Oren et al., 2023; Li and Flanigan, 2024). These issues are directly

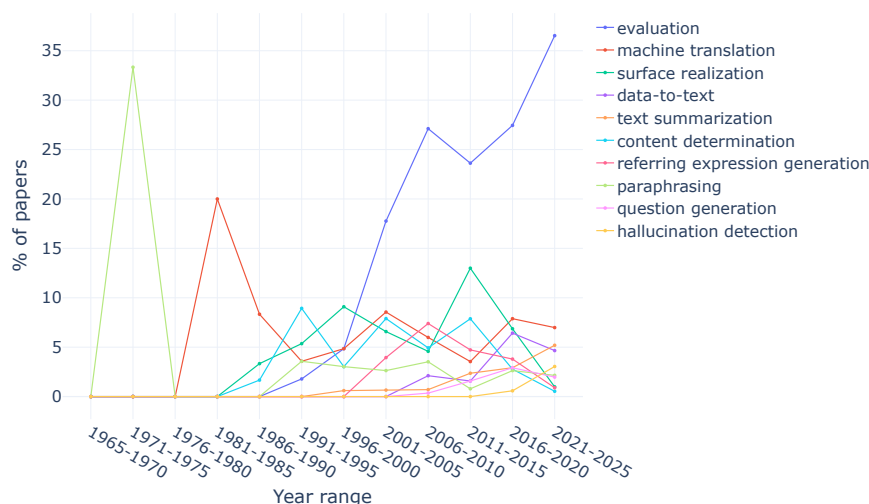


Figure 2: Fine-grained task trends in the NLG domain. Percentages indicate the share of papers addressing each of the overall ten most common tasks, across five-year periods.

relevant in the context of NLG, where, e.g., [Balloccu et al. \(2024\)](#) show that a number of GPT models have been globally exposed to millions of samples from hundreds of benchmarks. Closely related are privacy risks, where information from training data may be memorized and inadvertently leaked by models ([Yan et al., 2025](#)), with particularly serious implications in clinical domains ([Alber et al., 2025](#)). This concern is especially relevant for medical-related NLG applications, such as medical note generation ([Knoll et al., 2022](#); [Moramarco et al., 2022](#); [Savkov et al., 2022](#)), where generated texts may involve sensitive information. Furthermore, training data are likely to contain biased or stereotyped information regarding numerous demographic characteristics, which may lead to harmful biases in LLMs' outputs ([Resnik, 2025](#)). Specific work has been carried out in NLG to evaluate demographic bias in generated texts, with methodologies for such evaluations proposed in, e.g., [Aggarwal et al. \(2022\)](#).

With LLMs now constituting a quarter of NLG studies and only likely to increase, these multi-faceted concerns should raise awareness among researchers. We encourage researchers to both apply LLMs judiciously and be open to exploring approaches beyond LLMs.

5 Conclusion

There have been concerns within the NLP community about the increasingly inward-looking nature of NLP research. It has been noted that some strands of current research risk sidelining linguistic theory, even though it remains relevant in many aspects ([Opitz et al., 2025](#)). Relatedly, attention has also been drawn to the deas-

ing interaction with other language-related research communities ([Reiter, 2007b](#)), further reinforcing concerns about the narrowing scope of the field. Especially at a time when AI research is at an all-time high and questions are being raised about the sustainability of academic conferences ([Chen et al., 2025](#)), our work sought to understand whether, in the age of LLMs, NLG research has become inward-looking as well. So far, the use of LLMs for NLG research, while significant, does not constitute the totality of NLG research. Additionally, the arrival of these models has led to an increase in interest in tasks, which could be explained by the new opportunities for research presented by these models. Evaluation has seen greater focus of researchers' attention as the use of probabilistic language models poses new questions and challenges in understanding how to qualify performance and task fitness.

There is nevertheless a danger for NLG researchers that placing our limited attention on one particular language-generation paradigm may lead to a dead end or underplay the true extent of what is attainable for certain tasks, particularly as there is also a growing symbiotic relationship between NLP research and the few industry players that fund this research through grants and internships ([Abdalla et al., 2023](#)). Furthermore, there is the broader question of whether the current LLM paradigm of data and compute leads to low-resource languages or other tasks being neglected.

Our message to researchers is that NLG needs diversity of approaches and ideas beyond a singular paradigm. There are opportunities for intrepid researchers to take on problems or tasks that are seemingly neglected by others that would enrich us all.

Acknowledgements

We are grateful to Ondřej Dušek, Ehud Reiter, and the anonymous reviewers for constructive feedback on this work.

References

- Abdalla, Mohamed, Jan Philip Wahle, Terry Ruas, Aurélie Névél, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13141–13160, Toronto, Canada. Association for Computational Linguistics.
- Aggarwal, Arshiya, Jiao Sun, and Nanyun Peng. 2022. Towards robust NLG bias evaluation with syntactically-diverse prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6022–6032, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alber, Daniel Alexander, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A. Valliani, Jeff Zhang, Gabriel R. Rosenbaum, Ashley K. Amend-Thomas, David B. Kurland, Caroline M. Kremer, Alexander Eremiev, Bruck Negash, Daniel D. Wiggan, Michelle A. Nakatsuka, Karl L. Sangwon, Sean N. Neifert, Hammad A. Khan, Akshay Vinod Save, Adhith Palla, Eric A. Grin, Monika Hedman, Mustafa Nasir-Moin, Xujin Chris Liu, Lavender Yao Jiang, Michal A. Mankowski, Dorry L. Segev, Yindalon Aphinyanaphongs, Howard A. Riina, John G. Golfinos, Daniel A. Orringer, Douglas Kondziolka, and Eric Karl Oermann. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2):618–626.
- Amato, Sara and Kutz Arrieta. 2023. Natural language generation in the logos model. In *Proceedings of the 1st International Workshop on Multilingual, Multimodal and Multitask Language Generation*, pages 32–39, Tampere, Finland. European Association for Machine Translation.
- Augenstein, Isabelle, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.
- Azime, Israel Abebe, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadgign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, Assefa Atsbiha Tesfu, Philipp Slusallek, Tamar Solorio, and Dietrich Klakow. 2025. ProverbEval: Exploring LLM evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6265–6281, Albuquerque, New Mexico. Association for Computational Linguistics.
- Balloccu, Simone, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.
- Barzilay, Regina and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 331–338, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Basmov, Victoria, Yoav Goldberg, and Reut Tsarfaty. 2024. Simple linguistic inferences of large language models (llms): Blind spots and blinds.
- Belz, Anya, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.
- Belz, Anya, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Belz, Anya, Anastasia Shimorina, Maja Popović, and Ehud Reiter. 2022. The 2022 ReproGen shared task on reproducibility of evaluations in NLG: Overview and results. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 43–51, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

- Belz, Anya and Craig Thomson. 2023. The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Belz, Anya and Craig Thomson. 2024. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA and ICCL.
- Belz, Anya, Craig Thomson, Javier González Corbelle, and Malo Ruelle. 2025. The 2025 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 1002–1016, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Berglund, Lukas, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Stickland, Tomek Korbak, and Owain Evans. 2024. The reversal curse: Lms trained on “a is b” fail to learn “b is a”. In *International Conference on Learning Representations*, volume 2024, pages 18623–18642.
- Bouyamourn, Adam. 2023. Why LLMs hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3181–3193, Singapore. Association for Computational Linguistics.
- Branco, António, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.
- 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.
- Calderon, Nitay and Roi Reichart. 2025. On behalf of the stakeholders: Trends in NLP model interpretability in the era of LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 656–693, Albuquerque, New Mexico. Association for Computational Linguistics.
- Cambaz, Doga and Xiaoling Zhang. 2024. Use of AI-driven Code Generation Models in Teaching and Learning Programming: a Systematic Literature Review. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1, SIGCSE 2024*, pages 172–178, New York, NY, USA. Association for Computing Machinery.
- Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).
- Chen, Nuo, Moming Duan, Andre Huikai Lin, Qian Wang, Jiaying Wu, and Bingsheng He. 2025. Position: The current ai conference model is unsustainable! diagnosing the crisis of centralized ai conference.
- Cripwell, Liam, Anya Belz, Claire Gardent, Albert Gatt, Claudia Borg, Marthese Borg, John Judge, Michela Lorandi, Anna Nikiforovskaya, William Soto-Martinez, and Craig Thomson. 2023. The 2023 WebNLG shared task on low resource languages. overview and evaluation results (WebNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 55–66, Prague, Czech Republic. Association for Computational Linguistics.
- Dale, Robert. 2023. Navigating the text generation revolution: Traditional data-to-text nlg companies and the rise of chatgpt. *Natural Language Engineering*, 29(4):1188–1197.
- Dargis, Roberts, Guntis Bārzdīņš, Inguna Skadiņa, Normunds Grūzītis, and Baiba Saulīte. 2024. Evaluating open-source LLMs in low-resource languages: Insights from Latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293, Miami, USA. Association for Computational Linguistics.

- Das, Mayukh and Wolf-Tilo Balke. 2025. From prototypical to relational: How LLMs navigate complex analogies. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 465–485, Hanoi, Vietnam. Association for Computational Linguistics.
- van Deemter, Kees. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.
- Di Eugenio, Barbara, Michael Glass, and Michael Troilo. 2002. The DIAG experiments: Natural language generation for intelligent tutoring systems. In *Proceedings of the International Natural Language Generation Conference*, pages 120–127, Harriman, New York, USA. Association for Computational Linguistics.
- Diebold, Francis X. 2012. On the origin(s) and development of the term “big data”. Technical report, University of Pennsylvania, PIER Working Paper No. 12-037. Available at SSRN.
- Dong, Yihong, Xue Jiang, Jiaru Qian, Tian Wang, Kechi Zhang, Zhi Jin, and Ge Li. 2025. A Survey on Code Generation with LLM-based Agents. ArXiv:2508.00083 [cs].
- Dušek, Ondřej and Filip Jurčiček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. 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.
- Friedrich, Annemarie. 2025. Conference report acl 2025 - how can we move forward as a field? <https://annefried.github.io/posts/2025-08-02>. Accessed: 2026-01-26.
- Gala, Jay, Pranjal A Chitale, AK Raghavan, Varun Gumma, Sumanth Doddapaneni, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.
- Gatt, Albert and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61(1):65–170.
- Gatt, Albert, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the Neonatal Intensive Care Unit: Using NLG technology for decision support and information management. *AI Commun.*, 22(3):153–186.
- Gatt, Albert and Ehud Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93, Athens, Greece. Association for Computational Linguistics.
- Gehrmann, Sebastian, Falcon Dai, Henry Elder, and Alexander Rush. 2018. End-to-end content and plan selection for data-to-text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Gkatzia, Dimitra and Saad Mahamood. 2015. A snapshot of NLG evaluation practices 2005 - 2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60, Brighton, UK. Association for Computational Linguistics.
- González Corbelle, Javier, Alberto Bugarín-Diz, Jose Alonso-Moral, and Juan Taboada. 2022. Dealing with hallucination and omission in neural natural language generation: A use case on meteorology. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 121–130, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.
- Guhe, Markus. 2007. *Incremental conceptualization for language production*. Routledge: Taylor and Francis Group.
- Guo, Yuting, Anthony Ovadge, Mohammed Ali Al-Garadi, and Abeed Sarker. 2024. Evaluating large language models for health-related text classification tasks with public social media data. *Journal of the American Medical Informatics Association : JAMIA*, 31(10):2181–2189.
- Gururaja, Sireesh, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.
- Hagos, Desta Haileselassie, Rick Battle, and Danda B. Rawat. 2024. Recent Advances in Generative AI and

- Large Language Models: Current Status, Challenges, and Perspectives. *IEEE Transactions on Artificial Intelligence*, 5(12):5873–5893.
- Hedderich, Michael A., Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.
- Hoang, Andrew, Antoine Bosselut, Asli Celikyilmaz, and Yejin Choi. 2019. Efficient adaptation of pre-trained transformers for abstractive summarization.
- Howcroft, David M., Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Howcroft, David M. and Dimitra Gkatzia. 2022. Most NLG is low-resource: here’s what we can do about it. In *Proceedings of the Second Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 336–350, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.
- Hunter, James, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada, and Cindy Sykes. 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine*, 56(3):157–172.
- Issaka, Sheriff, Keyi Wang, Yinka Ajibola, Oluwatumininu Samuel-Ipaye, Zhaoyi Zhang, Nicte Aguilon Jimenez, Evans Kofi Agyei, Abraham Lin, Rohan Ramachandran, Sadick Abdul Mumin, Faith Nchifor, Mohammed Shuraim, Lieqi Liu, Erick Rosas Gonzalez, Sylvester Kpei, Jemimah Osei, Carlene Ajeneza, Persis Boateng, Prisca Adwoa Dufie Yeboah, and Saadia Gabriel. 2025. The African Languages Lab: A Collaborative Approach to Advancing Low-Resource African NLP. ArXiv:2510.05644 [cs].
- 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):248:1–248:38.
- Jurafsky, Daniel and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released August 24, 2025.
- Karamolegkou, Antonia, Sandrine Schiller Hansen, Ariadni Christopoulou, Filippos Stamatiou, Anne Lauscher, and Anders Søgaard. 2025. Ethical concern identification in NLP: A corpus of ACL Anthology ethics statements. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11618–11635, Albuquerque, New Mexico. Association for Computational Linguistics.
- Karanikolas, Nikitas, Eirini Manga, Nikoletta Samaridi, Vaios Stergiopoulos, Eleni Tousidou, and Michael Vassilakopoulos. 2025. Strengths and Weaknesses of LLM-Based and Rule-Based NLP Technologies and Their Potential Synergies. *Electronics*, 14(15):3064.
- Keskar, Nitish Shirish, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.
- Khan, Mohammed Safi Ur Rahman, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. IndicLLM-Suite: A Blueprint for Creating Pre-training and Fine-Tuning Datasets for Indian Languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.
- Knoll, Tom, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 385–394, Seattle, United States. Association for Computational Linguistics.
- Kostikova, Aida, Zhipin Wang, Deidamea Bajri, Ole Pütz, Benjamin Paaßen, and Steffen Eger. 2026.

- LLMs: A data-driven survey of evolving research on limitations of large language models. *ACM Comput. Surv.* Just Accepted.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Lee, Ji-Ung, Haritz Puerto, Betty van Aken, Yuki Arase, Jessica Zosa Forde, Leon Derczynski, Andreas Rücklé, Iryna Gurevych, Roy Schwartz, Emma Strubell, and Jesse Dodge. 2023. Surveying (dis)parities and concerns of compute hungry nlp research.
- Li, Changmao and Jeffrey Flanigan. 2024. Task Contamination: Language Models May Not Be Few-Shot Anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480.
- Liu, Peter J., Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.
- Liu, Xiao, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.
- Lyu, Mengxian, Xiaohan Li, Ziyi Chen, Jinqian Pan, Cheng Peng, Sankalp Talankar, and Yonghui Wu. 2026. Natural language generation in healthcare: A review of methods and applications. *Journal of Biomedical Informatics*, 176:104997.
- Mahamood, Saad and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21, Nancy, France. Association for Computational Linguistics.
- Maurya, Kaushal Kumar and Ekaterina Kochmar. 2025. Pedagogy-Driven Evaluation of Generative AI-Powered Intelligent Tutoring Systems. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, pages 32–46, Cham. Springer Nature Switzerland.
- Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- McCann, Bryan, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering.
- Mei, Hongyuan, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730, San Diego, California. Association for Computational Linguistics.
- Mille, Simon, Elaine Uí Dhonnchadha, Stamatia Dasiopoulou, Lauren Cassidy, Brian Davis, and Anya Belz. 2023. DCU/TCD-FORGe at WebNLG’23: Irish rules! (WegNLG 2023). In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.
- Min, Bonan, Hayley Ross, Elinor Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. ArXiv:2111.01243 [cs].
- Moramarco, Francesco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.
- Movva, Rajiv, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. 2024. Topics, authors, and institutions in large language model research: Trends from 17K arXiv papers. In *Proceedings of the 2024 Conference of the*

- North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1223–1243, Mexico City, Mexico. Association for Computational Linguistics.
- Mpagouli, Aikaterini and Ioannis Hatzilygeroudis. 2009. A Knowledge-based System for Translating FOL Formulas into NL Sentences. In Iliadis, Maglogiann, Tsoumakasis, Vlahavas, and Bramer, editors, *Artificial Intelligence Applications and Innovations III*, volume 296, pages 157–163. Springer US, Boston, MA. Series Title: IFIP Advances in Information and Communication Technology.
- Occhini, Giulia, Kumiko Tanaka-Ishii, Anna Barford, Refael Tikochinski, Songbo Hu, Roi Reichart, Yijie Zhou, Hannah Claus, Ulla Petti, Ivan Vulić, Ramit Debnath, and Anna Korhonen. 2026. Artificial intelligence is creating a new global linguistic hierarchy.
- Ojo, Jessica, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How Good are Large Language Models on African Languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.
- Opitz, Juri, Shira Wein, and Nathan Schneider. 2025. Natural Language Processing RELIES on Linguistics. *Computational Linguistics*, 51(3):1009–1032.
- Oren, Yonatan, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2023. Proving Test Set Contamination in Black-Box Language Models. In *The Twelfth International Conference on Learning Representations*.
- 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Portet, François, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789–816.
- Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. Association for Computational Linguistics.
- Qin, Libo, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2025. Large Language Models Meet NLP: A Survey. ArXiv:2405.12819 [cs].
- Ravichander, Abhilasha, Shruti Ghela, David Wadden, and Yejin Choi. 2025. HALoGEN: Fantastic LLM Hallucinations and Where to Find Them. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1402–1425, Vienna, Austria. Association for Computational Linguistics.
- Reiter, Ehud. 2007a. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 97–104, Saarbrücken, Germany. DFKI GmbH.
- Reiter, Ehud. 2007b. The shrinking horizons of computational linguistics. *Computational Linguistics*, 33(2):283–287.
- Reiter, Ehud. 2025a. *Natural Language Generation*. Springer Cham.
- Reiter, Ehud. 2025b. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.
- Reiter, Ehud. 2026. NLG evaluation: Past, present, future. In *Proceedings of the 1st Symposium on Natural Language Generation Evaluations*, pages 8–15, Aberdeen, United Kingdom. Association for Computational Linguistics.
- Reiter, Ehud and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.
- Reiter, Ehud, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1):137–169.
- Resnik, Philip. 2025. Large Language Models Are Biased Because They Are Large Language Models. *Computational Linguistics*, 51(3):885–906.
- Rogers, Anna. 2023. Closed ai models make bad baselines. <https://hackingsemantics.xyz/2023/closed-baselines>. Accessed: 2025-12-10.
- Rohatgi, Shaurya, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

- Same, Fahime, Guanyi Chen, and Kees Van Deemter. 2022. Non-neural models matter: a re-evaluation of neural referring expression generation systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5554–5567, Dublin, Ireland. Association for Computational Linguistics.
- Savkov, Aleksandar, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. Consultation checklists: Standardising the human evaluation of medical note generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120, Abu Dhabi, UAE. Association for Computational Linguistics.
- Schmidtova, Patricia, Eduardo Calò, Simone Balloccu, Dimitra Gkatzia, Rudali Huidrom, Mateusz Lango, Fahime Same, Vilém Zouhar, Saad Mahamood, and Ondrej Dusek. 2025. Do my eyes deceive me? a survey of human evaluations of hallucinations in NLG. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 60–79, Hanoi, Vietnam. Association for Computational Linguistics.
- Se, Ksenia. 2024. The recipe for an ai revolution: How imagenet, alexnet and gpus changed ai forever. <https://www.turingpost.com/p/cvhistory6>. Accessed: 2026-01-29.
- Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.
- Thomson, Craig and Ehud Reiter. 2020. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.
- Vatsal, Shubham, Harsh Dubey, and Aditi Singh. 2025. Multilingual Prompt Engineering in Large Language Models: A Survey Across NLP Tasks. ArXiv:2505.11665 [cs].
- Wen, Tsung-Hsien, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.
- Weston, Jason, Frédéric Ratle, and Ronan Collobert. 2008. Deep learning via semi-supervised embedding. In *Proceedings of the 25th international conference on Machine learning*, pages 1168–1175.
- Yan, Biwei, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2025. On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. *High-Confidence Computing*, 5(2):100300.
- Zubiaga, Arkaitz. 2024. Natural language processing in the era of large language models. *Frontiers in artificial intelligence*, 6:1350306.