Evaluating Task-oriented Dialogue Systems: A Systematic Review of Measures, Constructs and their Operationalisations

Anouck Braggaar, Tilburg University, the Netherlands A.R.Y.Braggaar@tilburguniversity.edu

Christine Liebrecht, Tilburg University, the Netherlands C.C.Liebrecht@tilburguniversity.edu

Emiel van Miltenburg, Tilburg University, the Netherlands C.W.J.vanMiltenburg@tilburguniversity.edu

Emiel Krahmer, Tilburg University, the Netherlands E.J.Krahmer@tilburguniversity.edu

Abstract This review gives an overview of evaluation methods for task-oriented dialogue systems, discussing the constructs, metrics and operationalisations used in previous work and highlighting the challenges in the context of dialogue system evaluation. The objective of this review is to encourage a more critical approach when evaluating dialogue systems. To that end, a systematic review of four databases was conducted (ACL, ACM, IEEE and Web of Science), which after screening resulted in 122 studies. Those studies were carefully analysed for the constructs and methods they proposed for evaluation. Four of the most occurring constructs (*satisfaction*, *correctness*, *quality*, and *efficiency*) are discussed as an example of how constructs are operationalised and measured in research. Additionally, recent developments regarding large language models are discussed for their applicability in the context of evaluation of dialogue systems. Furthermore, considerations and concerns about validity and reliability are discussed in relation to the found constructs and metrics. To improve consistency in evaluation approaches, future work should take a critical and systematic approach to the operationalisation and specification of the used constructs. To work towards this aim, this review ends with a research agenda for dialogue system evaluation and suggestions for outstanding questions.

1 Introduction

Over the last few years, dialogue systems (schematically depicted in Figure 1) have become much more robust, and practical applications are within reach and even in existence already. This development resulted in an increased interest in dialogue system research both in practice as well as in science (Følstad et al., 2021). The 2022 survey by Costello and LoDolce (2022) illustrates this increased interest in the application of dialogue systems, demonstrating that in the customer service domain, already 54% of the surveyed companies use a dialogue system when communicating with their customers. Additionally, Costello and LoDolce (2022) predict that by 2027, dialogue systems will be the primary channel for approximately a quarter of organisations. Although these systems are often employed in practice, users often feel frustrated with the interactions (Press, 2023). This frustration might even cause users to avoid using a dialogue system in the future. A recent survey shows that 30% of customers either abandon the brand or tell their friends about the bad experience with the customer service dialogue system (Press, 2023).

To avoid such bad experiences, it is important to accurately define and assess the capabilities of a dialogue system. Therefore, evaluation of such systems remains an important task. However, many challenges arise when trying to evaluate a dialogue system. In practice and in research many distinct evaluation metrics and constructs are being used, but it is not always clear what is being measured and how the evaluation is conducted. The selection of constructs and methods thus seems to be a challenging task. Moreover, it is hard to define what a *good* dialogue or dialogue system is and how this quality could eventually be captured in a measure (Deriu et al., 2021). There are many different constructs that can be measured; it depends on the task and context which of those are relevant.

As a result, evaluation needs to be done with great care to ensure reliability and reproducibility of the results. However, there seems to be a lack of standardisation, regarding both metrics, constructs and their operationalisations (as mentioned for example by Casas et al. 2020). Thus, proper evaluation of dialogue systems is important, as a well-functioning system is essential for both the user and the stakeholders behind the dialogue system (such as the organisation developing and

https://doi.org/10.3384/nejlt.2000-1533.2026.5940

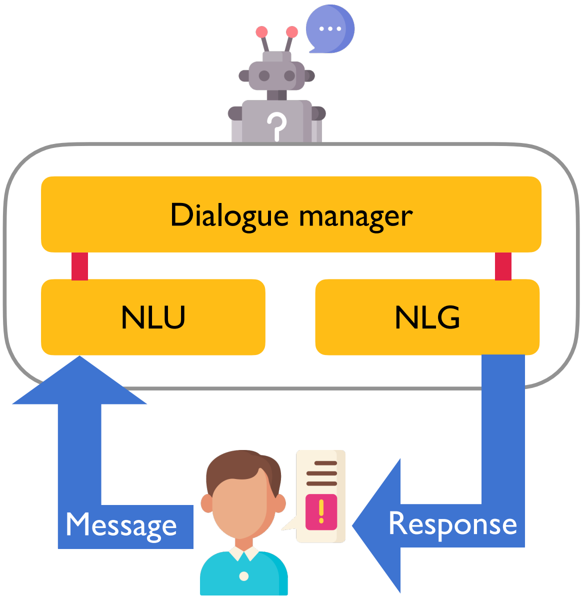


Figure 1: Simplified depiction of an interaction with a task-oriented dialogue system. Internally, messages are traditionally processed through a Natural Language Understanding (NLU) module, after which a dialogue manager updates the internal state of the system, and selects an appropriate response, which is then realised by the Natural Language Generation (NLG) module. (Icons via Freepik.com.)

implementing the system). To work towards this aim, this review will take a critical approach towards current evaluation procedures currently being used in literature.

1.1 Dialogue Systems

There is a high degree of variation in the terminology used to refer to task-oriented dialogue systems. While Jurafsky and Martin (2024) argue that chatbots are distinct from task-oriented dialogue systems as chatbots are designed for more unstructured conversations, in the literature the aforementioned terms tend to be used interchangeably. In this review, both the terms 'dialogue systems' and 'chatbot' will be used to refer to task-oriented systems.

ELIZA (Weizenbaum, 1966) is seen as the first chatbot to have been developed. It was based on a small set of rules and keywords, enabling it to respond to users with either a pre-programmed response or a variation of the user's own utterances. Although Weizenbaum has repeatedly stated that ELIZA was created as a *parody* of Rogerian psychotherapists,[1] others nonetheless took ELIZA quite seriously, as a first step towards automating psychological treatment (Weizenbaum 1976 cites Colby et al. 1966 as an example). More importantly, and again to Weizenbaum's surprisal, people started to have deep conversations with ELIZA, and were quick to anthropomorphise the system (Weizenbaum, 1976). This shows the impact that dialogue systems can have on users, even with relatively simple means. A final observation Weizenbaum (1976) made (perhaps connected to the anthropomorphisation of ELIZA) was "the spread of a belief that [ELIZA] demonstrated a general solution to the problem of computer understanding of natural language" (Weizenbaum, 1976, p. 7), even though this is demonstrably false. We currently see a similar kind of wishful thinking around the performance of Large Language Models (LLMs; Mitchell 2021). To us, these observations highlight the importance of both (i) critically thinking about what it *means* for a system to have particular cognitive/linguistic abilities, and (ii) the use and developments of valid and reliable evaluation methods that test the abilities that a system has, and the impact that the system has on the user.

This literature review focuses on task-oriented dialogue systems as illustrated in Figure 1. In interactions with task-oriented dialogue systems, the user first presents a problem that they would like to solve. Through a series of messages to and responses from the dialogue system, both interlocutors work towards finding a resolution. The processing and generating components shown in Figure 1 may vary from straightforward rule-based to more complex machine learning approaches (Harms et al., 2018). Different input methods can be used, all with different processing methods. Users can click on buttons (the dialogue system will then by necessity follow a straightforward predefined script) or make use of free input fields (which needs intent-recognition for handling the user's request). Nowadays, task-oriented dialogue systems can be much more complex than systems like ELIZA, that mostly relied on rule-based algorithms. While rule-based systems are unlikely to disappear as that they are easy to maintain and adjust, reliable and predictable (Leusmann et al., 2024; McTear and Ashurkina, 2024), LLMs are increasingly used for powering dialogue systems. A recent example involves the work by Chung et al. (2023), who use LLMs in their framework for creating an end-to-end task-oriented dialogue system. The advantages of LLMs are clear: the output is often more fluent, and the enormous amount of training data means they have a broad vocabulary and can produce rich and varied texts straight away. That said, there are still many open questions in the literature on the use of LLMs for dialogue systems. In our view, regardless of the technology that is used, in all cases evaluation is an important factor to take into account. Although new

[1] See Yao and Kabir (2023) for a brief introduction to Rogerian psychotherapy.

technologies might result in more and different evaluation metrics, the basic ideas behind evaluation remain the same.

1.2 Constructs and Measurement

Evaluating a dialogue system involves characterising the external behaviour of the system, its internal workings, or the effects that the system has on either its users or on other processes that the system is embedded in. For this characterisation, different ideas or concepts are employed to help explain the situation. For example, the high *readability* and *accuracy* of the generated responses might make a system *easy to use*, which increases the users' *efficiency* and *intention to use* the system again. Following longstanding tradition in psychology, we refer to these ideas or concepts as *constructs* (Cronbach and Meehl, 1955).[2] The term *construct* is also often used within the field of Natural Language Generation (NLG) when discussing evaluation (see for example Van der Lee et al. 2021). Since constructs are fairly abstract concepts, they are not directly measurable. To do so, they need to be *operationalised*, i.e. "define them in such a way that they can be measured" (Treadwell, 2017, p. 30). For example, recent work has already used these terms to define and measure model bias in NLP (Van der Wal et al., 2024). Figure 2 provides an illustration of this idea - multiple metrics can measure the same construct, but at the same time capture different aspects of the construct. Meaning that in some cases the construct might not be fully represented in the operationalisation. To summarise, we will use the following terminology in this review:

- *Construct:* Ideas or concepts to explain situations, e.g. user satisfaction.
- *Operationalisation:* Definition of the construct so that the construct can be measured, e.g. are the user's expectations met?
- *Metric/measure:* How the construct is actually measured, e.g. surveys or questionnaires.

As will become evident, research on dialogue systems has explored a variety of constructs, with different studies operationalising the same construct in many distinct ways[3]. This is problematic, as it is not always clear what exactly is being measured. In this review we encourage a more critical approach to the operationalisation of constructs and associated metrics. In our results section we will show the most used constructs in literature and the most used metrics to measure these constructs. Additionally we will also discuss the operationalisations of the constructs. To not overload the reader we will only focus on the discussion of four of the most used constructs as an example to show how one can reason about such constructs. Additional tables with all the found constructs can be found on OSF[4]. Through our construct-driven approach, we are also able to contrast different operationalisations of the same construct, showing how they each focus on different aspects of the ideas they aim to capture. Through our work, we provide a template for future researchers to critique and systematically approach the operationalisation of different constructs. This allows for a higher degree of reproducibility and a more valid comparison between studies.

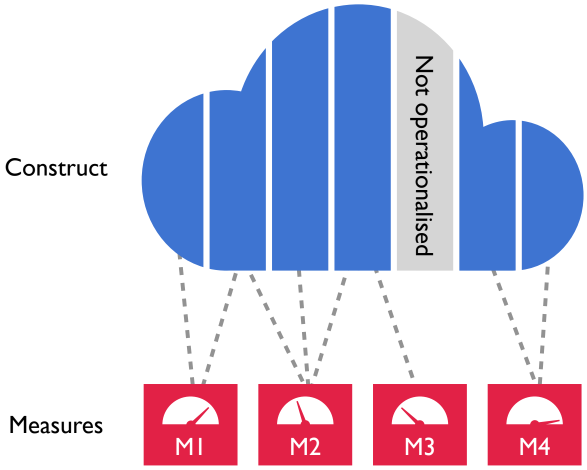


Figure 2: Different measures (M1…M4) operationalising the same construct, capturing different aspects. We may obtain a fairly good coverage of the construct by combining different metrics, but some aspects may remain elusive.

1.3 Why this Review?

Many reviews on dialogue system (evaluation) have been published, as Table 1 demonstrates. These reviews often have a different scope than the current review. There are reviews on dialogue systems in general, or specifically focusing on the question of evaluation of dialogue systems. Some reviews focus on specific technical aspects while other reviews narrow the scope by focusing on systems in a specific domain.

The current review therefore aims to provide a critical discussion of the different evaluation constructs and metrics for task-oriented *textual* dialogue systems. This paper serves two goals. First, we will show the vast

[2] Strauss and Smith (2009) provide an in-depth discussion of the origins and current debates around the idea of construct validity.

[3] Confusingly, different authors also (i) refer to the same constructs with different names, or (ii) refer to different constructs with the same names. This observation has also been made in NLG research by Howcroft et al. (2020).

[4] https://osf.io/x2qja/overview

	Evaluation	Task-oriented	Constructs
Federici et al. (2020)	+	+	−
Abd-Alrazaq et al. (2020)	+	+	−
Valizadeh and Parde (2022)	+	+	−
Deng et al. (2023)	+	+	−
Ni et al. (2023)	+	+	−
Algherairy and Ahmed (2024)	+	+	−
Yi et al. (2024)	+	+	−
Edwards and Mason (1988)	+	−	+
Abu Shawar and Atwell (2016)	+	−	+
Ren et al. (2019)	+	−	+
Casas et al. (2020)	+	−	+
Finch and Choi (2020)	+	−	+
Motger et al. (2022)	+	−	+
Liu et al. (2016)	+	−	−
Maroengsit et al. (2019)	+	−	−
Fan and Luo (2020)	+	−	−
Deriu et al. (2021)	+	−	−
Yeh et al. (2021)	+	−	−
Caldarini et al. (2022)	+	−	−
Nakano et al. (2025)	+	−	−
Singh and Namin (2025)	+	−	−
Jannach et al. (2021)	−	+	+
Peng and Ma (2019)	−	+	−
Syvänen and Valentini (2020)	−	+	−
Zhang et al. (2020b)	−	+	−
Deng et al. (2025)	−	+	−
Mariani et al. (2023)	−	−	+
Chen et al. (2024b)	−	−	+
Cui et al. (2020)	−	−	−
Kusal et al. (2022)	−	−	−
Park et al. (2022)	−	−	−
This paper	+	+	+

Table 1: The focus of previous overviews on (evaluation of) dialogue systems.

amount of different constructs and operationalisations and discuss four of the most used constructs in depth. Additionally, we will discuss how LLMs can be used for dialogue system evaluation and how these new developments relate to existing approaches. Furthermore, we will examine how the concepts of validity and reliability relate to evaluation approaches discussed in this paper. Second, we will show how evaluation can be approached while taking into account a broader perspective (such as the customer service domain). The paper ends with a research agenda that aims to stimulate follow-up research on the evaluation of task-based dialogue systems, and to generate mutual agreement on the different constructs, operationalisations and measures used for chatbot evaluation in general. Although developments in dialogue systems research are evolving rapidly, we believe that the constructs and metrics used will remain relevant over the coming years.

1.4 Reading Guide

This review will start with the Method (§2) in which we will show the way in which the literature was selected (§2.1 and §2.2) and how subsequently data was extracted (§2.3). Next we will discuss in Section 3 how one can reason about these constructs and propose a preliminary division into categories. In Section 3.1 the top-10 most occurring constructs in the literature are discussed together with the definitions of the top-four constructs. Then in Section 3.2 we will discuss how these four constructs can be operationalised and show the most common metrics in the literature. Then we will catch up with current work involving LLMs in the evaluation process (§3.3). In Section 4 the validity and reliability of evaluation measures will be discussed. Additionally, we will focus on triangulation and combining different evaluation approaches (§4.7) and discuss the current problems around standardisation (§4.8). Finally, the customer service domain will be used as a case study to show that it is always important to take the context into account when evaluating a system. Section 5 contains an example conversation and a step-by-step evaluation to illustrate how these findings can be applied in practice. This is followed by a discussion on the challenges of evaluation within the customer service domain (§6). We will end this review with the limitations (§7) and a brief conclusion (§8) proposing outstanding questions and recommendations.

2 Method

2.1 Databases and Search Queries

The first round of literature selection concerned the selection of databases that would be used for finding the literature. Four databases were chosen that contain published papers focusing on the more technical (NLP related) fields (ACM[5], ACL anthology[6], IEEE[7] and Web of Science[8]). The search needed to be as comprehen-

[5] https://dl.acm.org/search/advanced

[6] https://aclanthology.org/

[7] https://ieeexplore.ieee.org/search/advanced

[8] https://www.webofscience.com/wos/woscc/advanced-search

sive as possible so no time periods were specified and the default setting of the respective database-search engines were used. To make sure papers were included that mainly focused on dialogue systems and involved some kind of evaluation, only the title and abstract were searched (not the full text). In all databases the following search query was used:

(chatbot OR 'dialogue system*' OR 'dialog system*')*
AND
(eval OR analy* OR perf* OR perc*)*

The query ensured that plurals and different spellings of the words were selected. This means for example that papers containing the keywords *chatbots* and *evaluating* or *analysing* were selected but also articles on *dialogue systems*, *analyzing* and *performing* or *perceptions*. We chose to restrict to the terms dialogue system and chatbot because these are often used within NLP and can be used interchangeably.

For ACM the full text collection was searched. In IEEE, the default settings were used, and in Web of Science all editions of the Web of Science collection were searched. The searches were done on the eighth and ninth of December 2021. For ACL, the ACL anthology BibTeX download including abstracts (08-12-2021) was used. Our goal was to make an inventory of constructs and metrics - with the current time frame (up to and including 2021) we can make a thorough and complete assessment. The explosion of papers (see also Figure 4) makes a larger scan infeasible. Of course, major post-2021 developments such as the rise of LLMs requires additional attention. These developments will be discussed in Section 3.3.

In total 3,800 papers were found using this search strategy. From the results, duplicate entries were removed based on the title and DOIs. If the title and DOI were the same, only one entry was kept. If there were doubts (e.g. title is the same but DOI was different) these were marked and evaluated manually. This meant that eventually 3,458 records were kept for the first round of manual selection. Code and data for screening of the duplicates and the further selection process can be found on OSF.

2.2 Paper Selection

We used the PRISMA approach (Figure 3) to filter out irrelevant papers and to obtain a manageable subset for further analysis. We considered a paper to be relevant when there is some sort of (reflection on) evaluation of a task-oriented dialogue system.

A first quick selection consisted of screening the title and abstract by the first author. Based on automatically searching for keywords, papers were either retained or discarded. In case of doubt, the paper was retained in this stage of the procedure to make sure no papers were missed. Papers that included key phrases like spoken, interface, open-domain, emotion and annotation were thought to be about different sort of systems (such as social agents) or focus on different aspects of a dialogue system (such as the interface), and were discarded. Of the 3,458 papers screened, 2,807 papers were manually removed.

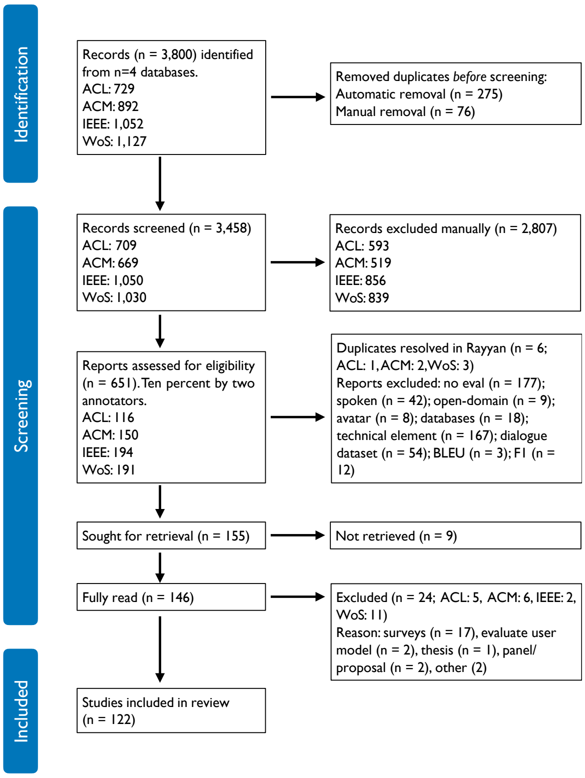

Identification
Records (n = 3,800) identified from n=4 databases.
ACL: 729
ACM: 892
IEEE: 1,052
WoS: 1,127
Removed duplicates *before* screening:
Automatic removal (n = 275)
Manual removal (n = 76)
Screening
Records screened (n = 3,458)
ACL: 709
ACM: 669
IEEE: 1,050
WoS: 1,030
Records excluded manually (n = 2,807)
ACL: 593
ACM: 519
IEEE: 856
WoS: 839
Reports assessed for eligibility (n = 651). Ten percent by two annotators.
ACL: 116
ACM: 150
IEEE: 194
WoS: 191
Duplicates resolved in Rayyan (n = 6; ACL: 1, ACM: 2, WoS: 3)
Reports excluded: no eval (n = 177); spoken (n = 42); open-domain (n = 9); avatar (n = 8); databases (n = 18); technical element (n = 167); dialogue dataset (n = 54); BLEU (n = 3); F1 (n = 12)
Sought for retrieval (n = 155)
Not retrieved (n = 9)
Fully read (n = 146)
Excluded (n = 24; ACL: 5, ACM: 6, IEEE: 2, WoS: 11)
Reason: surveys (n = 17), evaluate user model (n = 2), thesis (n = 1), panel/proposal (n = 2), other (2)
Included
Studies included in review (n = 122)

Figure 3: PRISMA figure showing the selection process.

To finalise the selection of papers relevant for the aim of this review, we used Rayyan (Ouzzani et al., 2016, a collaborative online platform for carrying out systematic reviews), to create and manage annotations. Using Rayyan, an extra round of duplicate removal was done, resulting in 645 papers for the next selection phase. By means of a flowchart with exclusion and inclusion criteria (see OSF), ten percent of the 645 papers were annotated by two annotators. Papers that for example focus on a virtual avatar were excluded from the data set. In addition, to narrow the selection further, papers that only use BLEU or F-scores for evaluation were also removed since a first screening showed that many papers use these metrics without further discussing evaluation and often only focusing on specific elements of dialogue systems. Without a proper reflection on evaluation, these standardised metrics become less meaningful. As will be evident from the results section of our review (§3), these metrics are still also present in our

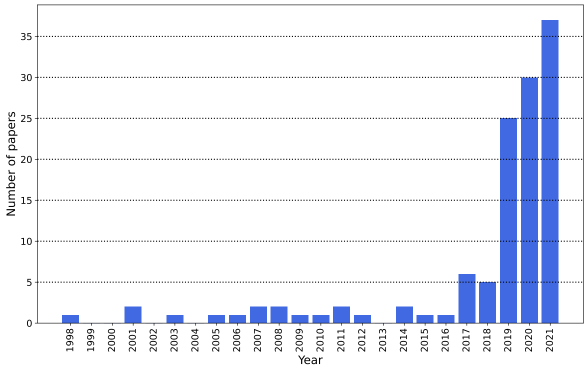


Figure 4: Bar graph showing the occurrence of papers within each year.

final selection since these metrics were also often used with other accompanying metrics in previous papers, or they were accompanied by a broader discussion of the evaluation method in these papers. This broader analysis fits better with the scope of the current review.

Ten papers were used for training the second annotator (an independent researcher who works with chatbots and who has less knowledge about technical aspects of evaluation). After training and discussion, the flowchart was accordingly updated and ten percent (i.e., 65 papers) of the sample was annotated by both annotators. Yet another paper was found to be duplicate, resulting in 64 doubly annotated papers. Papers were annotated based on a full read of the title and abstract. Out of 64 papers there were 13 disagreements on discard/retain. This resulted in a Kappa statistic of 0.491, indicating moderate agreement. The 13 disagreements were easily resolved after discussion, and the decision tree was updated to clarify a few remaining ambiguities. The remaining papers were assessed for eligibility using the decision tree. In total 155 records were kept, out of which nine could not be retrieved.

Out of the resulting 146 papers, 17 papers turned out to be reviews or overviews themselves (included in Table 1). These were excluded from the study as they themselves sum up metrics (which might result in double counts of constructs in our results). Another two papers were excluded because they focused specifically on evaluation of the user model instead of a dialogue system itself. One addition was a thesis, which was excluded because there was a similar paper of the author included in the selection. Two reports were excluded because they were a panel summary and a proposal. Another two were excluded because they were either unclear about what is being measured or how it should be measured.

Subsequently, a total of 122 papers were included in the current review. All of these papers were included, even if they turned out to not completely focus on task-oriented textual systems because these papers still can provide valuable insights in the evaluation of dialogue systems. As can be seen in Figure 4 there is a steep slope in how many papers are being published in this domain, with an increase especially since 2019. Since carrying out a systematic review at this scale is time-consuming, many new papers are published after the initial selection period. Therefore, we manually looked trough the literature after 2021 to ensure that recent developments have also been taken into account. We will reflect on these developments in Section 3.3.

2.3 Data Extraction Sheet

A data extraction sheet was made to systematically record relevant information from each paper. As a first step this sheet was piloted on ten of the included studies to make sure all important and interesting information was included. The data extraction sheet was divided into general information and study specific information. Table 2 shows the outline of the sheet.

Bibliographical and contextual information of the papers was first recorded. This included information like the title but also the domain and goal of the papers. Next, information on evaluation and measurement was recorded. Metrics were documented together with information on if they were specific for dialogue system evaluation and if they needed a reference for comparison. Then the construct and evaluators were recorded. It was documented if human and/or automatic evaluation was used. If human evaluation was used it was recorded who was the one evaluating (authors, experts, participants, users). Finally, some system details and additional information was documented. It was documented if the authors provided a critical analysis of the evaluation method, if they included statistics on the evaluation and if there was any qualitative analysis on the evaluation outcomes.

We tried to fill in the sheet as much as possible but in some cases this was not always feasible as some papers were quite vague and did not always mention all (sometimes important) details. This complicated our comparison of the different papers, and also made it very challenging to reproduce the experiments. This is a well-known issue often discussed within the field of NLG and NLP (see for example Howcroft et al. 2020; Belz et al. 2023). We will further reflect on this in the section about validity and reliability (§4).

3 Constructs and Measurement

In the 122 papers, 109 distinct constructs were found that were used for evaluation. There are many ways to examine these constructs and the associated metrics. Distinctions between constructs are somewhat ar-

Bibliographical	**Measurement data**	**System details**
Title	Metric	Type of system
Author(s)	Construct	Goal/purpose of system
Year	Evaluator	Language
Journal		Implementation of system
Language		
Context	**Evaluation details**	**Additional information**
Domain or industry	Data set used	Critical analysis
Goal of paper	Sample size	Statistics on evaluation
	Turn or conversation level	Qualitative analysis
	Moment of evaluation	Reflection on difficulty
	Intrinsic/extrinsic	Other comments

Table 2: Information extracted from papers.

bitrary and it is entirely possible that some constructs fit into multiple categories. Table 3 shows a preliminary distinction into two perspectives: intrinsic and extrinsic. This division is merely a way to create a simpler overview over the 109 found constructs. The complete overview of these categories and constructs along with the metrics can be found on OSF.

Operationalising constructs and measuring relevant variables can provide insights on properties of the system on its own. This is often referred to as intrinsic evaluation (see for example Resnik and Lin 2010). Within intrinsic evaluation we identified three subcategories (Table 3). The first subcategory, **natural language understanding (NLU)**, focuses on *understanding* the users' utterances. This can be considered as a key factor for a well-functioning dialogue system. After all, if the system fails at this part, the utterance is likely to be misunderstood and the chances are high that the user will not be content with the response and hence the overall system. As can be seen in Table 3, we could only associate two constructs with NLU. With just ten papers mentioning constructs related to NLU, it seems fair to say that less attention is devoted to this category than to the other categories related to intrinsic evaluation. The second subcategory, **natural language generation (NLG)**, focuses on the texts that are *produced* by the system. These constructs are typically operationalised using metrics that are also used in the NLG field (see Celikyilmaz et al. 2020, for an extensive overview). The last subcategory focuses on **performance and efficiency** related constructs, often focusing on the question of how (efficiently) the system performed in general or in certain tasks, taking into account aspects like time or costs.

Secondly, constructs can provide insights about a system in a certain context. This is often referred to as extrinsic evaluation (Resnik and Lin, 2010). We also identified three subcategories within extrinsic evaluation. Firstly, **task success and effectiveness** includes constructs that focus mainly on the outcomes of a task or intervention. Not surprisingly, this subcategory contains a number of constructs that can be explicitly linked to certain domains and contexts. For example, comparing treatments and health outcomes are of importance in a healthcare context, knowledge of material and learning outcomes for education, and (customer) loyalty is key for a customer service chatbot. Next we identified the subcategory **usability** which focuses on constructs like the ease of use, accessibility and learnability. Usability is often also seen as a construct of its own, which can be defined as: "the capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and user support, to fulfil the specified range of tasks, within the specified range of environmental scenarios" (Shackel, 2009, p. 340). Lastly, **user experience** was identified as a subcategory. This category contains constructs related to users' perceptions of the system. Where usability mostly focuses on the practical evaluation of a system, user experience more closely incorporates the perceptions of the users. User experience contains the most identified (distinct) constructs, highlighting the interest in the perceptions of the users with regards to their usage of the systems.

In the end, it can be a puzzle to categorise the constructs as it can be difficult to see how they relate to each other and there can be many different definitions for a single construct (see also §4.8). Therefore, Table 3 does not show a final categorization but a possible way to categorise constructs. In the next sections we will therefore not focus on all of these categories, but will discuss in detail the constructs that are most often found in the data. By exploring four example constructs, we hope to show how one can reason about these constructs and their measurement. In the next section (§ 3.1) we will first show and discuss *what* is being measured in the context of task-oriented dialogue system evaluation. Next, we will discuss *how* these con-

Perspective	Subcategory	Number of constructs	Examples
Intrinsic evaluation	Natural language understanding	2	Context-capturing, understanding
	Natural language generation	31	Coherence, naturalness, relevance
	Performance/efficiency	15	Compatibility, efficiency, robustness
Extrinsic evaluation	(Task) success /effectiveness	15	Effectiveness, health outcomes, (customer) loyalty
	Usability	9	Ease of use, intention to use, learnability
	User experience	37	Engagement, enjoyment, satisfaction

Table 3: Number of constructs (total of 109) and examples categorised by perspectives and categories.

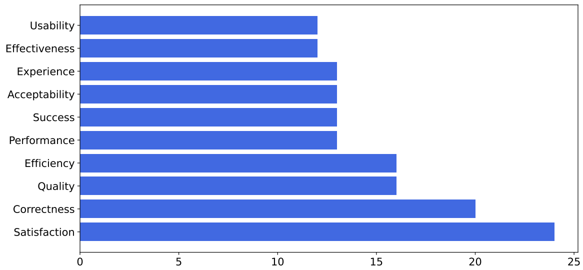


Figure 5: Top-10 most occurring constructs in dialogue system evaluation research.

structs are actually being measured (§ 3.2). Finally, we will examine the role of LLMs in the context of dialogue system evaluation (§ 3.3).

3.1 What is being Measured?

Seventy-two constructs appear at least twice in our data, with 31 of them occurring at least five times. Figure 5 shows the ten most measured constructs. Both satisfaction and correctness occur 20 times or more, quality and efficiency are used 16 times, and all other constructs occur less than 15 times. We will focus on the four constructs most frequently used (*satisfaction, correctness, quality, and efficiency*). *Satisfaction* (extrinsic evaluation) occurred most often, while the other three most common constructs (*correctness, quality, and efficiency*) focus on intrinsic evaluation. This shows that researchers have a strong focus on measuring the capabilities of a system on its own, but also frequently take the user's perspective into account.

Satisfaction

Most often, researchers focus on the construct satisfaction. Satisfaction reflects if users are satisfied and content with the system and/or conversation. In some cases, this might also involve how satisfied users are with the outcomes of the conversation (such as reaching their goal). User satisfaction can consist of several components as described by Maier et al. (1997): "the user gets the information s/he wants, is comfortable with the system, gets the information in an acceptable elapsed time" (Maier et al., 1997, p. 9). Research often follows previous work on (user) satisfaction for defining this construct. For example, Eren (2021) follow the *expectation-confirmation theory* (ECT) from Oliver (1980) to define user satisfaction. According to this theory, satisfaction is defined by whether perceived performance of the system aligns with (previous) user expectations (Oliver, 1980).[9] Abu Shawar and Atwell (2007) follow the definition given by Maier et al. (1997) in which user satisfaction is partly expressed in terms of perceived usefulness and usability. Satisfaction is a complex and multifaceted construct, this is why in these definitions it is often defined in terms of other constructs. As we will see in Section 3.2, this also means that we cannot operationalise this construct through a single question. Instead, researchers should ask users multiple questions to accurately capture their level of satisfaction with the system.

Correctness

Correctness tends to focus on how accurate, correct and precise the utterances of a dialogue system are. Some of the papers in the selection address the state of the given information, evaluating if the information returned by the system is factual and correct (Campillos-Llanos et al., 2020, 2021). Su et al. (2020) use the term *factuality* to refer to the veracity of the output. Veracity of a text can be reinforced by other constructs, e.g. as a text is deemed very fluent one might think the chances are higher that the text is correct (Van Deemter, 2024). Sometimes researchers thus connect correctness to a

[9] For a more detailed discussion on the expectation-confirmation theory see §6.2.

different construct, such as relevancy. For example, Duggenpudi et al. (2019) ask users to rate the system output using the question/description "How relevant/correct is the answer retrieved." This illustrates that there are many different ways to interpret correctness. Involving multiple constructs into the question asked to the user might be confusing, therefore researchers should both carefully consider how the constructs should be defined and how this is reflected in the question (Van der Lee et al., 2021).

Quality

The third most used construct for dialogue system evaluation is quality. Defining quality seems to be quite difficult, because what does it actually mean to be of high quality, and how does it relate to other constructs like correctness? There are many papers discussing quality as an autonomous construct. Finch et al. (2021), for example, aim to measure overall dialogue quality but do mention that the interpretation of dialogue quality is especially hard for chat-oriented systems. As an example, they mention that misunderstandings might indicate low quality. It could be argued that the same also holds for task-oriented systems. Quality can also be applied to more specific parts of the dialogue system, such as the quality of information, as for example in Gonzales and González (2017), or the generation quality as discussed in Shi et al. (2021). Some of the usages of quality raise the question as to how this construct is different from other constructs.

The quality of information as used by Gonzales and González (2017) mirrors the definition of correctness, as this quality might also refer to the factuality of the utterances (among other constructs). Consequently, quality seems, in most cases, to incorporate multiple different constructs.

The difficulty of defining quality actually arises from the fact that the definition of quality sometimes consist out of multiple different constructs. Shi et al. (2021) for example involve other constructs to measure the quality of responses, namely *nonrepetiveness*, *consistency*, and *fluency*. The quality of the responses is thus measured by combining multiple constructs that are also used separately in the literature. This shows that quality is a complex construct, often defined in terms of other constructs that seem to be somewhat subjectively selected.

Efficiency

Efficiency often seems to be defined as accomplishing a specific goal given, for example, a certain time frame or other specified 'costs'. Often research refers to the PARADISE framework when defining and operationalising efficiency (especially when they refer to the 'costs') (Walker et al., 1997). The PARADISE framework was specifically designed for spoken dialogue systems. The goal of the framework is to minimise the costs to eventually reach user satisfaction. The costs can be measured by efficiency measures such as time - the costs can therefore refer to anything that you *don't* want from your system (Walker et al., 1997). For example, in the context of task-oriented systems the system should probably be quick in answering the user query (e.g. defined in the number of turns or the time taken for the complete conversation). Bickmore and Giorgino (2006) for example measure efficiency as part of the 'costs' defined by PARADISE, eventually resulting in one overall quality score. Similarly, Foster et al. (2009) create three categories based on PARADISE - dialogue efficiency, dialogue quality and task success. Dialogue efficiency in their cases also focuses on time (measuring time taken, mean time of the system to respond and number of turns).

The costs seem to be specific to the task that needs to be accomplished or the function of the dialogue system (e.g. Takanobu et al. 2020 define efficiency in terms of accomplishing the task of the task-oriented system). In cases of task-oriented systems, for example, the users often want quick interactions, therefore minimizing the time seems to be efficient. On the other hand, for open-domain or chit-chat systems, users may expect longer interactions. Therefore, researchers need to consider their specific context to define efficiency. Thus, efficiency is task and context dependent.

Conclusion. Overall, there are many constructs that are used in research to evaluate dialogue systems but in some cases one could wonder if they are not too vague or if constructs are used interchangeably (see also the discussion in Howcroft et al. 2020 and Fitrianie et al. 2020). A construct like *quality* is very broad and not all researchers define what they mean with the construct (which makes comparison more challenging). Similarly, a construct like *correctness* might be used as part of *quality*. It makes one wonder how informative a measurement of the construct is. These complications highlight the importance for researchers to clearly define the construct of interest, motivate why this construct is of importance and carefully consider the operationalisation.

3.2 How are Constructs Measured?

In this section we will describe how the constructs are operationalised, while illustrating how they are actually measured in research. In general, a division can be made between automatic and human evaluation approaches. While automatic approaches often focus on intrinsic evaluation, human evaluation aims to capture

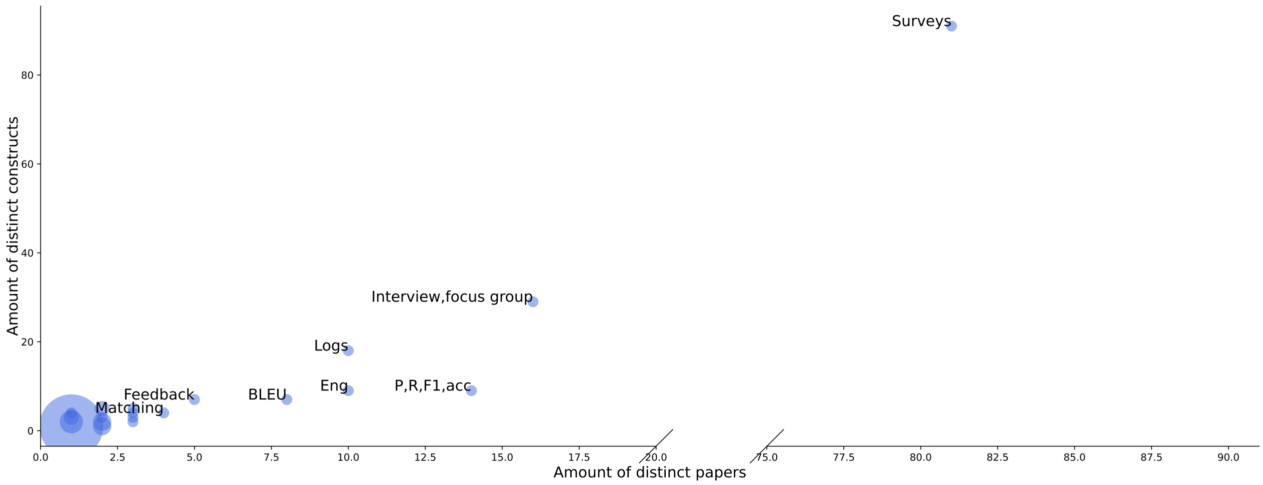


Figure 6: Amounts of distinct constructs and papers associated with approach. Bubble size indicate frequency. (Eng - engagement metrics; P,R,F1,acc - precision, recall, F-score and accuracy)

how users actually perceive the system (extrinsic evaluation; see §3 for a more detailed explanation on intrinsic versus extrinsic evaluation). This section will begin with a first impression based on the results in Figure 6. Following this, the same constructs as in Section 3.1 (top four most measured constructs) will be used as examples to illustrate how constructs can be operationalised when evaluating a dialogue system.

A first impression. Figure 6 shows the number of distinct papers and constructs associated with each approach. Surveys and ratings (asking users to rate statements without having a full survey) are by far the most used metrics (categorised together as ‘surveys’). This approach was used in 81 papers and with 91 different constructs. A reason for why surveys are popular is probably because they can easily be reused from previous work and can be optimized according to the researchers own needs.

The figure shows that in general many human approaches are used, such as interviews, focus groups, and obtaining feedback. These human approaches offer a high degree of flexibility in what they can measure, so it is logical that they measure many different constructs and are also used in many distinct papers. Some of the approaches are in-between human and automatic measures such as analysing logs or using engagement metrics. Analysing logs can be done manually but also automated. Engagement metrics (such as duration of the conversation or the length of the responses) can also be measured using both human approaches as well as automatically. With these metrics, researchers attempt to objectively measure the users’ attitude (in terms of the construct ‘engagement’ but also in terms of other constructs such as ‘intention to use’).

The automatic metrics are often adopted from other domains within NLP such as machine translation. The figure shows that metrics like precision, recall and BLEU (and variants of BLEU such as delexicalized-BLEU or kn-BLEU) are often used (even though we filtered out papers that exclusively evaluate using f-score/recall/precision metrics, see §2.2). A more extensive discussion on these metrics will be provided in the discussion on *Correctness* and *Quality*. More surprising is the variety of different constructs that are measured by an approach like BLEU. The BLEU metric, as defined by Papineni et al. (2002), compares a system utterance to a reference-utterance and therefore seems to determine the quality of an utterance based on the similarity to a reference. It is surprising, then, that over five different constructs are measured by BLEU.

Lastly, in the lower-left corner of Figure 6 a big bubble indicates that there are many metrics that are only used to measure a few constructs by a handful of papers. This includes for example metrics that have been introduced by these papers, often specifically intended for dialogue system evaluation. We will discuss some of the approaches in more detail below by taking the four most used constructs (as outlined in §3.1) as examples. More information on how these constructs are measured can be found on OSF.

Satisfaction

Satisfaction is a multi-faceted construct as Section 3.1 already showed. This construct is sometimes considered a hard to quantify, all-compassing construct as it covers all aspects of a system, from recognition to generating output (Aust and Ney, 1998). Although the definition may seem broad, the construct is in practice mostly measured by one approach: surveys. This is not surprising as satisfaction encompasses users' perceptions. Furthermore, surveys offer the advantage of incorporating multiple items to highlight the different aspects of satisfaction. The definition of satisfaction should match the operationalisation in the surveys. However, some work bases the operationalisation of satisfaction on only one construct and often ask participants to simply rate their satisfaction on a scale. For example, Kataoka et al. (2021) ask participants to rate on a 0 (highly dissatisfied) to 5 (highly satisfied) scale how satisfied they are with the chatbot. One could wonder if this question captures all aspects of the construct. As satisfaction is a multi-faceted construct, a single item probably does not fully capture this construct and multiple items might be the preferred option. Some of the authors actually do employ multiple items. Sensuse et al. (2019) for example utilize four items to measure user satisfaction.

While surveys are often used in research, the type of survey and items included can vary. Some of the used surveys are existing surveys that have been reused in the context of dialogue systems such as the System Usability Scale (SUS) (Brooke, 1996) and the Questionnaire for User Interface Satisfaction (QUIS) (Chin et al., 1988). The SUS-scale was introduced as a broad and quick measure for usability across multiple use-cases. It consists out of ten items that are rated on a five-point scale, such as "I thought the system was easy to use" (Brooke, 1996). Usability itself is a broader construct and Brooke (1996) argues that this construct consists of the sub-dimensions effectiveness, efficiency and satisfaction. In the literature, this scale is therefore often used to measure system satisfaction. The QUIS was developed in the context of software evaluation and focuses, as the name mentions, specifically on the interface of a system. It covers items such as "Organization of information on screen" (10 point scale from *confusing* to *very clear*) and "Task can be performed in a straightforward manner" (10 point scale from *never* to *always*). This survey seems to be somewhat broader as it also encompasses items that focus on system reliability and terminology used by the system.

Next to using existing surveys, researchers often develop their own surveys or refine surveys and base this on previous work as well. For example, Eren (2021) use items by Chung et al. (2020) who also base their items on previous work. Satisfaction is defined by the extent to which expectations are met (this definition was also discussed in §3.1) (Chung et al., 2020). Six items are used to measure satisfaction, including questions like "The service agent did what I expected" and "I am satisfied with the service agent" (Chung et al., 2020).

Only two other methods are used to measure satisfaction: informal user feedback (used once) and comparing to another system (also used once). In the first case users are prompted to provide general feedback on the system, in the latter case users are being asked to compare different versions of dialogue systems on this construct.

Overall, satisfaction is measured only by human approaches, primarily surveys. This is reasonable as this construct focuses on how users perceive the system. However, it is important to use multiple items to measure satisfaction as it is a multi-faceted construct. Additionally, researchers should carefully consider if there are pre-existing surveys that can be used and if these are suitable to measure satisfaction.

Correctness

In contrast to satisfaction, correctness is often measured using multiple automatic metrics. For example, accuracy (correct items divided by all items) is used by several papers, in some cases together with precision (number of items among selected items that are correct), recall (number of correct items that are correctly selected) and F1 (harmonic mean of precision and recall) (Jurafsky and Martin, 2024). With these metrics, researchers often focus on specific parts of the dialogue system such as intent recognition (the correctness of the predictions). Some of the metrics are specifically designed for the context of dialogue systems. An example of this is DialTest (Liu et al., 2021), which measures the accuracy of the intent recognition and robustness of the dialogue system. DialTest generates similar test cases and is able to select data that might trigger errors in the model. The data set is then used to test the robustness of the model or for retraining (Liu et al., 2021).

In addition to automatic metrics, human evaluation of correctness was carried out by means of a survey, analysing the logs, or by ranking. Campillos-Llanos et al. (2021), for example, analyse the conversation logs manually to verify if the information provided is correct compared to the information in the user record. When surveys are employed, sometimes only one item is used to ask participants for rating correctness (as we have seen in §3.1 with the example of Duggenpudi et al. 2019). Eric et al. (2017), for example, ask participants to evaluate output on correctness on a scale from one to five. Presumably, they use only one question to measure this. This assumes that there is one unambiguous way to measure correctness. This can be problematic as this is often not realistic for most constructs (see Howard 1981, on the *mono-operation bias*). All papers in our review actually employ only one item to measure correctness. This problem also does not mean that it is always necessary to use multiple items, but mono-operation brings the risk of only capturing a part of the construct. On the other hand, it can be expensive to incorporate multiple items, creating a trade-off between validity and efficiency.

Overall, automatic metrics provide an objective and clear picture of correctness, often focusing on specific elements of a dialogue system such as the intent recognition module. In contrast, human evaluation does not necessarily focus on one specific component of the system and can take a broader approach to correctness (taking into account for example contextual factors). Combining human and automatic approaches therefore could be a good venue for future evaluation attempts of correctness.

Quality

Quality can be measured both automatically and by human evaluation. Some of the automatic metrics are more straightforward and measure, for example, the response time and response length. Other metrics are adopted from other disciplines within NLP, such as machine learning. An example thereof is BLEU (used for example by Yin et al. 2017). The BLEU metric is an n-gram based textual similarity score (Papineni et al., 2002). This metric requires there to be a set of reference utterances to which an automatically generated utterance (the candidate) can be compared. Intuitively, BLEU looks at the overlap between the candidate and the reference utterances (in the case of Yin et al. 2017 the rewritten response is compared to the given query). This overlap is computed using the exact tokens in the sentence, meaning the BLEU score does not take synonyms into account, unlike alternative metrics such as METEOR (Banerjee and Lavie, 2005). The BLEU metric has extensively been used in machine translation and NLG, because for any given input, there is often only a limited set of possible translations or other appropriate outputs that need to be taken into account. Comparing automatically generated outputs to expected outputs makes sense, because a greater similarity could be expected between output and reference data to correlate with higher quality output. However, the BLEU metric has been criticised for its lack of correlation with human judgements, see for example Reiter (2018). Moreover, in the context of dialogue, one might wonder whether BLEU is still the right choice, since there are many possible responses to a given input.

There are also various different human evaluation approaches that can be used to measure quality. Most of the times this is either done by conducting a survey or asking for a rating. Measures can also be combined. Gonzales and González (2017) employ four different human approaches of measuring quality (specifically the quality of the information provided by the bot), namely conducting a survey, conducting an interview, directly observing participants and analysis of the usage policies of a (news providing) chatbot. Sensuse et al. (2019) focus on information quality, but also on service and system quality, following the information system success model by DeLone and McLean (1992). These different dimensions of quality can also encompass multiple other constructs (such as conciseness for information quality). The researchers measure these quality aspects by conducting a survey with items such as “ELISA can understand what I was asking for”.

In conclusion, quality can be assessed using both human and automatic metrics. The complexity of defining quality is reflected in the variety of different metrics and also in how these metrics are operationalised (e.g. multiple different items in a survey focusing on different aspects of quality). Combining different approaches could possibly lead to a more comprehensive assessment of the construct, considering both a quantitative evaluation as well as a more nuanced human judgement of quality.

Efficiency

The most straightforward way to measure efficiency is simply measuring the number of dialogue turns or time. This is for example done by Takanobu et al. (2020). Measurement of efficiency is most of the time (semi-)automatic or focuses on human evaluation. A measurement like time can both focus on ‘actual’ time as well as perceived time, since there can be a discrepancy between the perceived time and the actual time (as demonstrated by Thompson et al. 1996; perceived waiting time is more predictive for satisfaction than actual time). This raises the question if perceived efficiency is more important than ‘actual’ efficiency - the focus of the papers on human evaluation seems to suggest so. These perceptions are often examined through surveys. Roque et al. (2021) for example apply a standardised survey, namely the System Usability Scale (SUS, Brooke 1996; which was also used for satisfaction) which encompasses questions like “I found the system unnecessarily complex” and “I would imagine that most people would learn to use this system very quickly” (Brooke, 1996). Similarly, Pricilla et al. (2018) also ask users about the efficiency in a questionnaire, where participants had to rate on a five-point scale. With this questionnaire they measure multiple constructs next to efficiency such as effectiveness and helpfulness. However, it is not clear how these constructs are measured and how the questionnaire is developed.

Overall, efficiency can thus be measured (semi-)automatically or by human evaluation. However, there can be a discrepancy between perceived efficiency and ‘actual’ efficiency. Therefore, a combination of these approaches could potentially be the most informative to measure efficiency.

Conclusion. This section has shown that there are many different ways to measure evaluation constructs.

Notably, most of these metrics rely on human evaluation - particularly through surveys. However, the relation between these metrics and the definitions given by researchers is not always clear. Additionally, constructs are sometimes measured with only one survey item, raising concerns about the mono-operation bias. In some cases, survey items are not even reported clearly, which creates difficulties when comparing studies. Combining human and automatic metrics may provide a more comprehensive assessment of the constructs.

3.3 Using Large Language Models to Evaluate Dialogue Systems

Looking through the history of dialogue systems, it is clear that different kinds of technology may require different forms of evaluation to address their particular strengths and weaknesses. For example, research on hallucinations only became widespread after LLMs were adopted (see Ji et al. 2023 for a survey on hallucination). We believe that many of the current constructs will stay relevant over the next years, but there might be a shift in importance: a construct like fluency might be less relevant while a construct like factuality or correctness might gain importance when evaluating LLM based dialogue systems.

At the same time, LLMs may introduce a new dimension in automatic evaluation of dialogue systems. LLMs have great potential for the evaluation of dialogue systems, although there are also concerns. Previous work has already examined the use of LLMs in the context of NLG-evaluation (Li et al., 2024; Riyadh and Shafiq, 2023). In the context of dialogue systems, we believe LLMs can be used roughly in four ways:

1. To say something directly about the quality of a text, for example by means of a perplexity score to gauge the fluency of a text.
2. To train a regression model that predicts human quality scores either from the generated text alone[10] or from the generated text and a reference text. Examples include BLEURT (see Sellam et al. 2020, who augment BERT with pretraining on synthetic data. They test their model on the constructs fluency, grammar and semantics), BERTscore (see Zhang et al. 2020a, who compute similarity between a source sentence and candidate sentence) and COMET and its extension CometKiwi (see Rei et al. 2020, 2022, who also incorporate source language input).
3. To use the model instead of a human annotator. Through prompting the model, evaluations can be elicited. An example of this is GEMBA (Kocmi and Federmann, 2023), which shows state-of-the-art results in the field of translation quality assessment through zero-shot prompting (asking without further training whether the model can do something). Another example comes from Zheng et al. (2023), who discuss three types of *LLM-as-a-judge*: pairwise comparison (which response is the best), single answer grading (immediately giving a score), and reference-guided grading (using a reference answer to compare the output with). Recently, a shared task has been introduced to focus on prompting LLMs for evaluation of machine translation and summarisation (Leiter et al., 2023). Pradhan and Todi (2023) create in the context of this task five prompts to evaluate summarisation. They focus on prompting an overall score, coherence, consistency, fluency and relevance. Similarly, Akkasi et al. (2023) also participate in this task and prompt the models to evaluate coherence, completeness, conciseness, consistency, readability, syntax and a combination of all constructs.
4. To simulate user interactions with the dialogue system. Instead of relying on real users for evaluation, LLMs are sometimes employed to simulate users. This is often seen as a cost-effective strategy (De Wit, 2024). Researchers have used for example generative user simulators for reinforcement learning in task-oriented dialogue systems focusing on multi-domain goal state-tracking (Liu et al., 2022). ChatGPT has also been used to create simulated users for the evaluation of rule-based conversations (De Wit, 2024). Often these LLM based simulators are evaluated on user goal fulfilment and compared to either other simulators or human interactions (Davidson et al., 2023; Sekulić et al., 2024). Work by Meyer et al. (2022) examines if 'real' user data can be replaced by synthetic data generated by LLMs. In their zero-shot approach they prompt GPT-3 by asking questions in the domain of motivational interviewing through a conversational agent. They evaluate the performance of the synthetic data on a classification task (predicting three labels related to health changes) using a BERT-model with either original data, synthetic data, mixed data and mixed data with labels classified with a confidence level of 95% (Meyer et al., 2022).

The four ways described above all have their pros and cons. Many of the challenges that emerge when employing LLMs for evaluation concern validity and reliability. Reliability is generally good if a number of

[10] Note that this has a long history in machine translation (see e.g. Specia et al. 2018).

conditions are met to ensure that the system (both dialogue systems and evaluation models) is deterministic. In other words: it always gives the same output with the same input. Despite automatic solutions being deterministic, LLM-based systems can be brittle, with minimal changes to the input leading to different results. For example, previous work has shown that the order in which input is given changes the results of the task, showing a bias in ChatGPT for the first input item (Wang et al., 2023). To ensure validity, we must ask ourselves to what extent the model is able to 'capture' the relevant construct and to what extent this depends on the domain on which the model has been trained. Validity of LLMs will be discussed in more detail in Section 4.5.

4 Validity and Reliability

As noted in the introduction (§1.2), this paper concerns constructs and measurement; given a particular construct of interest, how can researchers operationalise that construct and actually measure to what extent a dialogue system is *satisfactory, correct, high-quality, efficient*, or …? To make all of this work, we need a deep understanding of these concepts, and how they relate to other concepts that we are interested in. Or at least: such an understanding is needed if we are to develop any kind of theory about how to build a good dialogue system that helps us achieve our goals. Furthermore, if we are interested to learn more about cognitive aspects of dialogue, the need for such an understanding is self-evident. This brings us to the question of validity. This section discusses some of the basics of validity theory.

Textbooks on research methodology (e.g. Bryman 2012; Treadwell 2017) often discuss validity in tandem with reliability. Generally speaking, *validity* is about measuring what you want to measure, and *reliability* is about the consistency of your measurements. Ideally, metrics should be both valid *and* reliable, since each is useless without the other; we cannot draw any conclusions from measures that are either meaningless or that deviate wildly from their intended target. In practice, there is often a trade-off between validity and reliability, since human ratings more closely match our experience (and are thus more valid), but they are more subjective (and thus less reliable) than automatic metrics. Automatic metrics are seen as offering quick heuristics or simplified proxies to the human experience (making them less valid), but they do provide consistent results (making them more reliable). Recent work in NLP by Van der Wal et al. (2024) discuss validity (in particular construct validity) and reliability and show how these perspectives can help improve the measurement of model bias.

There is a vast body of literature on the topic of validity (see the recommended readings in Fried and Flake 2018), but for brevity's sake we will focus on the 'Four Validities,' as presented by Vazire et al. (2022): construct validity (§4.1), internal validity (§4.2), external validity (§4.3), and statistical-conclusion validity (§4.4).[11] We shall only cover a selection of the issues that arise when looking at validity. In Section 4.5 we will then discuss the validity of LLM generated scores and in Section 4.6 we will discuss work on validity in the NLP field. Lastly, we will get back to the trade-off between human and automatic metrics (and when to use which kind of evaluation) in Section 4.7.[12]

4.1 Construct Validity

Construct validity "refers to the validity of inferences about how the measured or manipulated variables relate to the constructs of interest" (Vazire et al., 2022, p.163). First and foremost, authors should clearly define their construct of interest, so it is clear what is meant with the specific construct, and so that readers can assess the extent to which their measures operationalise that construct. As mentioned in our results section, few authors actually provided a definition. Moreover, where authors did define their constructs of interest, we found that different authors provided different (and sometimes incompatible) definitions for the same terms. This terminological confusion makes it hard to compare different papers.

Second, authors should provide enough information about how they operationalised the relevant constructs. Without this information, we also cannot tell whether their quality measures serve their intended purpose. To their credit, many authors do provide the code for their experiments (see Schmidtova et al. 2024 for a discussion on the availability of code in NLG), which in theory makes it possible to find out how they actually measured the quality of their models. However, we would still have to reconstruct the reasoning behind their approach, which is challenging to say the least. Furthermore, for human rating studies, it is absolutely essential to have a full specification of the experimental set-up. Without it, it is impossible to assess the construct validity of the study.

Third, authors can consider the evidence for the validity of their metrics. As an example, what evidence do we have that a Likert scale item such as 'this response sounds fluent' covers the full spectrum of what it means to be fluent? How do we know that participants' ideas

[11] The authors cite Shadish et al. (2002) as a source for this distinction, see their page 37 for the original definitions. Chapters 2 and 3 discuss their taxonomy of validities in more detail.

[12] The idea of reliability may also be tied to the idea of reproducibility (i.e., how repeatable are measures performed by different researchers?), but providing a full discussion goes beyond the scope of this review (see Belz et al. 2021 for an overview).

of fluency correspond to any established notion of fluency? And since different authors use different questions to assess the same constructs, what effect do all of these different formulations have on the outcomes of our experiments? The answer is that we do not know, and that hardly any papers provide any evidence for the validity of their metrics. Worse, still, despite the evidence against the validity of automatic quality measures such as the BLEU metric (e.g., Ananthakrishnan et al. 2007; Novikova et al. 2017; Sulem et al. 2018; Reiter 2018), these measures are still in use. Some papers in this review already spent some time discussing the validity of their automatic metrics. Both D'Haro et al. (2019) and Ye et al. (2021) propose a new automatic metric and show how their new metrics correlates to human evaluation.

4.2 Internal Validity

Internal validity refers to "the validity of causal inferences: Are assumptions upon which causal inferences are based explicitly stated and justified? Have plausible alternative explanations been convincingly ruled out?" (Vazire et al., 2022, p.164). Generally speaking, most inferences about dialogue systems in the NLP literature are fairly limited; the main goal seems to be to determine whether the proposed system is better than the alternative(s). Thus, the key independent variable is *System*, and the dependent variable is the quality metric of interest. The question, then, is whether the former has any impact on the latter. For reasons of space, we have not looked into the different system comparisons in detail, but in our experience the main threats to the internal validity in NLP are:

1. Confounding variables: when researchers present a new system and compare it to the state-of-the-art, we cannot know exactly what caused any differences in performance if the authors changed multiple variables at the same time.
2. Order effects: when participants always see the same items in the same order, this could potentially lead to a bias in their ratings (e.g. due to fatigue, or anchoring effects where the first few items serve as a reference point for the rest of the evaluation).
3. Lack of anonymisation: when participants know which system is which, this could potentially lead to them providing socially desirable responses (trying to please the researchers), rather than accurate assessments of system quality.

For a more in-depth discussion of potential issues in the design of human evaluations, we refer to Van der Lee et al. (2021).

4.3 External Validity

External validity refers to "the validity of inferences about how the observed effect will generalise beyond the specific conditions of the study" (Vazire et al., 2022, p.165). Given the characterisation of most NLP research, the conditions of most NLP studies could be defined in terms of three main components: participants, system properties, and context. Authors should make it clear to what extent they expect their findings to generalise towards other settings that differ in one or more of these dimensions. The following topics are pertinent to our discussion:

Sampling and score averaging. One question we may ask ourselves, for example, is whether the conversations with the system during the evaluation are representative for all possible conversations with the system. In this light, Van Miltenburg et al. (2021a) provide a discussion of different ways to sample the output-to-be-evaluated for human rating tasks or manual error analysis. Another question, often noted by Ehud Reiter (2017, 2022), is to what extent average-case performance is a good proxy for the user experience. Worst-case performance may be a better indication of the perceived quality of the system during real-world usage, since full breakdowns (however rare they may be) may render the system fully unusable.

Ecological validity. One important sub-category of external validity is *ecological validity*, which we might define as the extent to which the system, experiment, or metrics are true to reality. One question we can ask here is: to what extent is the system able to handle real-world conversations? (As opposed to conversations held in an artificial setting.) And what does it *mean* to hold a real-world conversation? Dingemanse and Liesenfeld (2022) discuss the importance of linguistically diverse conversational corpora to study these questions, and in follow-up work propose an evaluation approach to match their ambitions (Liesenfeld and Dingemanse, 2024).[13]

Reporting standards and design. Authors should also report all relevant characteristics of the sample, system, and context for us to assess any claims about generalisability. Based on earlier findings by Howcroft et al. (2020), we know that this is not the case: particularly human evaluations are often underdocumented.

Finally, authors should ensure that their claimed implications for future research or real-life applications are supported by the design of their study. This again

[13] Note that ecological validity does *not* require that dialogue systems themselves should appear humanlike, but they should be able to converse with humans. (Anthropomorphism is a contentious issue; see Abercrombie et al. 2023 for discussion.)

means that they should think carefully about the role that their constructs of interest play in a broader theoretical framework. As noted above, this requires clarity about the way those constructs are defined, and how they (supposedly) interrelate.

4.4 Statistical-Conclusion Validity

Statistical-conclusion validity refers to “the validity of statistical inferences” (Vazire et al., 2022, p.165). In recent years, this topic has started to receive more attention in NLP and NLG (e.g., Dror et al. 2018; Van der Lee et al. 2021; Van Miltenburg et al. 2021b), though the *reproducibility crisis* has put Psychology at the forefront of research on this topic. Since this goes beyond the scope of this review, we refer readers to the paper by Vazire et al. (2022) that we started this discussion with.

4.5 Validity of LLM-generated Scores and Simulated Users

In the field of machine translation, a lot of work has been done on (the validity of) quality estimation (Fonseca et al., 2019; Han et al., 2021). Recently, this kind of work has also focused on the possibilities of using LLMs for quality estimation (Huang et al., 2023). One of the problems with validity is the question of what aspects of the measured constructs are actually captured by the LLM in the evaluation. Some of the regression-based models are already able to model human predictions and achieve high correlations to these human scores, such as BLEURT (Sellam et al., 2020) and COMET (Rei et al., 2020). In the cases of BLEURT, research has shown that pre-training improves the robustness of these models in the case of domain shifts (Sellam et al., 2020). These regression based models are also compared to prompt-based methods. For example, Leiter et al. (2023) discuss the results of the shared task for prompting LLMs as metrics and for example compare newly created models to models such as BERTScore (Zhang et al., 2020a) in a machine translation context. In this case, one of the newly created models based on prompting actually achieves higher correlations to human scores than the baseline models such as CometKiwi (Rei et al., 2023) and BERTScore. Recent research examines prompting more closely and compares the outcomes of different prompts. In the context of the medical domain, Wang et al. (2024) show that multiple models behave different when given different prompt types (ranging from direct instructions to prompts that involve backtracking). The type of the prompt thus seems to matter to get reliable and consistent results.

Furthermore, Hu et al. (2024) show that LLMs actually confuse evaluation criteria (i.e. the constructs). For certain constructs, the scores generated by the model actually have a higher correlation with human ratings for another construct than with the human ratings for the indented construct. This is the case for example for LLM generated fluency scores, which show a higher correlation to human coherence scores than to human fluency scores. The authors also conclude that confusion issues shown by these models cannot be overcome by more elaborate definitions of a construct, while humans actually behave differently when given more elaborate definitions (Hu et al., 2024). Another question is how these models behave in different domains. Li et al. (2024) discuss the relevance of developing domain-aware models, as current models are often not specifically designed for one domain. This makes it difficult for those models to properly evaluate content in a specific domain (as for example a certain construct such as (medical) correctness is more important in a medical domain than in a customer service setting). Newly developed LLMs used for evaluation in specific domains should be made aware of domain-specific quality needs and constructs (Li et al., 2024).

Validity of using simulated users LLM-based simulations of users are often not seen as a replacement for real human evaluation (De Wit, 2024), raising the question of how these models actually reflect real user behaviour. Several studies have investigated if and how language models can model human behaviour in multiple domains. Argyle et al. (2023) use LLMs as a proxy for human populations in the context of for example vote prediction. Similarly, Horton (2023) uses LLMs to emulate experiments in an economical context. They both argue that this approach seems promising but also briefly discuss the negative consequences that these models can bring, such as dependency on owners of the models and misinformation (Argyle et al., 2023; Horton, 2023). Additionally, the work by Meyer et al. (2022) examined if real data (motivational interviews with a conversational agent) could be replaced by LLM generated data. The authors conclude that a classifier trained on synthetic data cannot reach the same performance as a classifier trained on real user data (showing also differences in language variability).

Domain dependency and societal inequality To train LLMs, quality filters are usually used to ensure that the model itself also generates language of acceptable quality. In a recent study, Gururangan et al. (2022) show that the GPT-3 quality filter is not neutral, but prefers texts from richer, urban and (on average) higher educated areas. So there is a confounding variable: while we would like to see that a language model can assess a text on a specific inherent quality dimension (such as fluency), it may be the case that the author of the text (who should not be relevant for our judgement) has an improper influence on the final score. Chen et al. (2024a) investigate the biases (such as an authority bias - having higher confidence in experts) present in LLM judges and human judges, and show that the five investigated biases are present in both human and LLM judges. Humans are not always outperforming the systems on certain biases showing that researchers should be aware of these problems both with LLM evaluation as well as with human evaluation (Chen et al., 2024a).

The dual role of LLMs LLMs can be used in multiple different roles and contexts, ranging from designing to evaluation. Previous work has done exactly this by employing ChatGPT as both the designer as well as the user that evaluates the product (Kocaballi, 2023). However, it is currently unclear what the implications are of using LLMs to both power *and* evaluate dialogue systems at the same time. A general rule in Machine Learning (ML) is that you should not test an ML system on training data, because then we will not get a good idea of the ability of systems to generalise to new data. But with LLMs used both in a dialogue system (to generate dialogue) and in the evaluation of that system (to evaluate the dialogue), it is unclear how generalisable the results of the evaluation are. A complicating factor here is that it is not always clear which data has been used as training data —LLMs use too much data to be able to document them afterwards (Bender et al., 2021).

4.6 NLP and Validity

Our discussion of validity is not to suggest that validity is not discussed at all in the NLP community, actually it is becoming more common. Below is a brief overview of relevant contributions.

Different studies in NLP (e.g., Kocmi et al. 2021; Moramarco et al. 2022) compare different metrics and human ratings, to see where they differ and where they agree. This is an example of convergent validity: testing whether measures that should in theory be related, are actually related. Xiao et al. (2023) go beyond correlation analyses, and provide an introduction to reliability and validity from the perspective of measurement theory, and translate these ideas into a set of tools (called *MetricEval*) to perform statistical analyses of NLG evaluation metrics. Our work in this paper is complementary to *MetricEval*, in that we take a more conceptual, high-level approach to validity.

Some recent papers also reflect on the status of benchmarks in our field. For example, Sun et al. (2023) test the concurrent validity of different benchmarks that aim to test compositional generalisation in LLMs, and shows that the use of different data sets results in a different model ranking. In a more theoretical paper, Schlangen (2021) presents an analysis of the way benchmarks are currently used to measure progress in our field. He argues that we should give more thought to the relation between data sets, tasks, individual cognitive capabilities, and overall language competence. In Schlangen (2021)'s view, we need to (re-)establish the connection between NLP tasks and related fields that study human competence in those areas. The paper by Sugawara et al. (2021) seems to do exactly this. They analyse the task of Machine Reading Comprehension (MRC; similar to NLU) from a psychological/psychometric perspective. The scholars follow Messick (1995) in distinguishing six aspects of validity[14] and translate these to the domain of MRC. In doing so, they establish what machine reading comprehension entails, and how to evaluate it. Subramonian et al. (2023) provide further reflections on the topic of benchmarks, based on a meta-analysis of the literature and a survey among NLP practitioners.

Finally, Sugawara and Tsugita (2023) discuss degrees of freedom in the way that researchers define and test systems for Natural Language Understanding (NLU). The authors provide a checklist for ensuring the validity of the validity of a test/benchmark. Although the paper is targeted at NLU, its arguments generalise to other areas of NLP.

We are happy that the question of validity is starting to receive more attention in the NLP literature, and fully support this movement. We recommend using the tools and checklists mentioned above, or the online Seaboat.io checklist, developed by Schiavone et al. (2023) which was also used as a guide for writing this section.[15]

4.7 Triangulation

This survey has provided an overview of different measures to quantify the performance of task-oriented dialogue systems. Here we will reflect on the type of approach; i.e., *how you want to measure a construct.* To be clear: there is no single best way to study the performance of a task-oriented dialogue system. Different approaches have different strengths and weaknesses, and every metric just highlights a subset of all possible quality dimensions. To fully understand the performance of a task-oriented dialogue system, it is necessary to combine different approaches. This is a practice known as *triangulation* (e.g., Noble and Heale 2019; Thurmond 2001). Thurmond (2001, p.253) defines it as "the combination of two or more data sources, investigators, [methodological] approaches, theoretical per-

[14] Referred to as *content, substantive, structural, generalizability, external*, and *consequential.* See the paper for definitions and more details.

[15] Alternatively, one might also refer to the checklist from Flake and Fried (2020), which they developed to avoid 'Questionable Measurement Practices.'

spectives (Denzin, 1970; Kimchi et al., 1991), or analytical methods (Kimchi et al., 1991) within the same study." We will first give a brief overview of the different approaches we have seen, and then consider ways to combine these approaches.

Asking People The first set of approaches used to evaluate task-oriented dialogue systems is to ask people about their interactions with the system. Human evaluation is generally still seen as the gold standard in NLG research, since automatic metrics are currently still unable to interpret and contextualise textual output as well as humans (Van der Lee et al., 2021). In theory, this could either be done before, during, or after interacting with the system. Although very few of the studies we looked at asked participants any questions *before* interacting with the dialogue system, this could be useful to gauge their expectations and perhaps their first impression of the system as has been shown in several studies in the fields of communication science (e.g., Van der Goot et al. 2021) and human-computer interaction (e.g., Khadpe et al. 2020).

We saw several studies that asked participants to reflect on their experiences *during* the interactions, for example using the *concurrent think aloud* study protocol (Holmes et al., 2019).[16] The authors describe this as a situation where the participant was recorded (both audio and video) while completing the task, talking through their observations and actions. This affords us more insight into the participants' experiences and thinking process. One important caveat, however, is that not all mental processes can be (accurately) verbalised. For further discussion of the origins and limitations of the think aloud protocol, see Nielsen et al. (2002). Finally, Fan et al. (2020) discuss how think aloud studies are generally used by UX-practitioners.

Most approaches are suitable for consultation *after* interacting with a dialogue system. Many studies opted for a survey, which was either developed by the authors themselves, or based on existing models, such as UTAUT-2 (Venkatesh et al., 2012) or the NASA Task Load Index (Hart and Staveland, 1988). While a self-authored survey does offer maximal flexibility, it hampers the ability to compare results between different papers. Hence we would recommend using pre-existing items (i.e., individual questions), scales (i.e. combinations of items measuring the same construct),[17] or questionnaires (i.e., combinations of scales that together provide an overview of relevant variables).[18]

Next to surveys, we also saw researchers carrying out interviews with individual participants and focus groups where a moderator leads a group discussion between participants, talking about their experiences. These methods are more qualitative in nature, and thus allow for a richer, more contextualised understanding of the participants' experiences. Interviews and focus groups are more common for Human-Computer Interaction researchers than in the Natural Language Processing community. For those new to these methods, we recommend the introduction by Bryman (2012).

Expert feedback can be gathered through having an expert either interact with the dialogue system itself, or having the expert look at (recorded) interactions between users and the system.

Looking at the Data Instead of asking people, we can also look at the interaction data ourselves. Although there may be a large amount of manual labour involved, there is no replacement to seeing what is going in the data. There are different kinds of data that may be used.

Most forms of evaluation involve spoken or written data: either a human or an artificial agent interacts with the dialogue system, and the interactions between them can simply be logged. These interactions also come with metadata, such as the number of conversational turns, response length, time spent on the task, and so on. Some scholars would consider these metadata as automatic metrics. As an example, Miraj et al. (2021) use different methods to measure both feasibility and acceptability, one of them being user engagement metrics. Similarly, Piau et al. (2019) use multiple metrics such as drop-out rates and the average time to answer questions to measure the construct acceptability.

An alternative to using spoken or written data is to go *beyond* these forms of communication. If the dialogue system is used by human participants, a video recording for example also allows analysis of their non-verbal responses (Holmes et al. 2019 used for example video and audio to record user experience), and the responses of others witnessing the interaction. Besides video, it may also be possible to capture biometric data of the participants. This is for example done by Przegalinska et al. (2019), who take psychophysiology metrics to measure the construct trust. Adding these approaches can create a more comprehensive overview of the users' perceptions.

Automatic Metrics The main benefit of automatic metrics is that they do not require any human assessments, making them quick and cost-effective (Deriu et al., 2021). Moreover, they are usually repeatable, and

[16] There are different kinds of think aloud protocols, but Alhadreti and Mayhew (2018) found that the concurrent think aloud method seems to give the best results.

[17] Scales often combine multiple items because it is often hard to capture a construct using a single item, and because it may be more reliable to average scores across multiple different items.

[18] Authors using crowdsourcing should also try to avoid the common mistakes described by Karpinska et al. (2021).

often come with a precise definition. The latter property also makes it possible to reason about their faithfulness in terms of the construct of interest. In the section about LLMs (Section 3.3) we have also discussed the advantages and disadvantages of using LLMs as a metric. We have shown that there are still issues with LLMs concerning reliability and validity of the generated scores.

Scholars should select relevant metrics based on their construct of interest, so that the evaluation scores are pertinent to their research question. Having that said, the speed and cost-effectiveness of automatic metrics mean that there is a relatively low threshold to publish more information about a dialogue system than is strictly necessary. For transparency reasons, scholars may wish to publish a larger set of scores to capture the overall performance of the system, and so that the scores can be scrutinised in future research.[19]

Combining Approaches There are two ways in which we can combine different approaches. Researchers and developers may use different evaluation approaches over time, or they synchronously use multiple evaluation strategies.

Evaluation starts when a project begins, and ends with the final assessment of the finished system. Different kinds of evaluation may be appropriate for different stages of the development process. By talking to experts and relevant stakeholders (users, developers, product owners, conversational designers etcetera) scholars develop a clear set of goals and key performance indicators for the dialogue system. During development, automatic metrics can be used to measure relevant variables directly, and to serve as proxies for variables that can later be measured more reliably by collecting user ratings and feedback. Once the (first version of the) dialogue system is ready, a more extensive evaluation can be carried out. At this stage one might also look at downstream effects on user behaviour and other business processes.

It may also be useful to use different evaluation strategies at the same time, since different evaluation approaches lead to different perspectives on the performance of a system. Again, this holds in two ways:

1. As we have seen in the results section, different metrics may capture different constructs. Thus, the quality of a dialogue system cannot be captured in a single number, which is why researchers need to be clear about the constructs of interest. If these are not specified *and* defined, it is unclear what the results even mean.
2. Constructs themselves may be complex or *multidimensional*. We have also seen this earlier: when multiple metrics operationalise the same construct, they may capture different aspects of what it means to be fluent, for example.[20]

Next to these evaluation strategies, one might also use multiple different quality measurement approaches to demonstrate *criterion validity*. For example, an automatic metric can be shown to have a high *concurrent validity* (a sub-type of criterion validity) if it shows a high correlation with human ratings of the same construct; for many metrics this has simply not been done yet. Demonstrating high concurrent validity is very useful if you have a larger project where it may not be feasible to run human evaluations for all experiments. If an automatic metric has a high correlation with human ratings for data in a particular domain, other researchers will probably have more confidence in the results obtained solely using the automatic measure.

4.8 The Need for Standardisation

As we have seen above, there is a wide range of different constructs that different researchers aim to measure. There is also a high degree of variation both in the terms used to refer to these constructs, as well as in the ways to operationalise them. As Howcroft et al. (2020) have also observed for human evaluation studies in the field of NLG: researchers may either use the same terms to refer to different constructs, or the other way round. This terminological confusion makes it hard to compare different results. Moreover, many studies fail to provide a definition for the constructs that are operationalised through their evaluation metrics, while some do not even mention the constructs of interest. Readers are left to wonder: *what is measured, exactly?*

Following Howcroft et al. (2020), we believe that it is important to standardise our evaluation terminology, and to improve our reporting standards. Some earlier studies that have worked towards standardisation are Belz et al. (2020) and Fitrianie et al. (2020). For human evaluation in particular, Shimorina and Belz (2022) provide a useful datasheet to include with any publication using human judges.

Beyond the standardisation of evaluation measures, there is also value in sharing model outputs in similar formats. For a concrete example, the GEM benchmark uses a common evaluation framework to facilitate model comparisons not just using existing metrics, but also using future metrics that are yet to be developed (Gehrmann et al., 2021, 2022).[21]

[19] This is reminiscent of the approach taken by the GEM benchmark (Gehrmann et al., 2021), where submissions are assessed with as many metrics as possible, to enable future system comparisons with relatively little effort.

[20] For further reading, Bryman (2012, Chapter 7) provides a useful discussion on multiple-indicator measures and multidimensionality —an idea he associates with the work of Lazarsfeld (1958).

[21] Earlier, Sedoc et al. (2018, 2019) presented a similar evaluation

5 How to Apply These Findings? A Case Study of Customer Service Chatbot Evaluation

The findings of this paper show what and how researchers measure when evaluating dialogue systems. However, the question remains how this knowledge can be applied in practice. In this section we will therefore discuss a fictitious conversation, shown in Figure 7, and demonstrate how readers can approach the evaluation of such a conversation/system. The example presents a conversation with both a customer service chatbot and a customer service employee. Imagine that this chatbot uses intent-recognition and pre-scripted responses. The customer wishes to return shoes. At first attempt the chatbot is unable to understand the user utterances and asks for a rephrase. The second attempt is understood, however the chatbot cannot deal with these kinds of questions yet and hands over to a human employee. Eventually, the employee is able to help the customer with the request. Three constructs will be discussed as an example for a possible evaluation approach, as these constructs differ in the way they are typically measured: fluency (mostly measured automatically), effectiveness (measured by human evaluation but also semi-automatically), and lastly satisfaction (measured only by human evaluation). To determine how to evaluate the conversation, we propose the following steps:

1. Determine the constructs of interest and their definition (in the case of the following example these are fluency, effectiveness, and satisfaction).

2. How is the construct typically evaluated, and what is the intuition behind these methods? Are there existing methods or standardized approaches that can be used to make the evaluation comparable to previous work?

3. How does the operationalisation of the construct relate to the case to be evaluated and its context?

4. Is this evaluation approach meaningful, feasible and worthwhile?

5. If this approach is not meaningful, is there an alternative to obtain an meaningful evaluation?

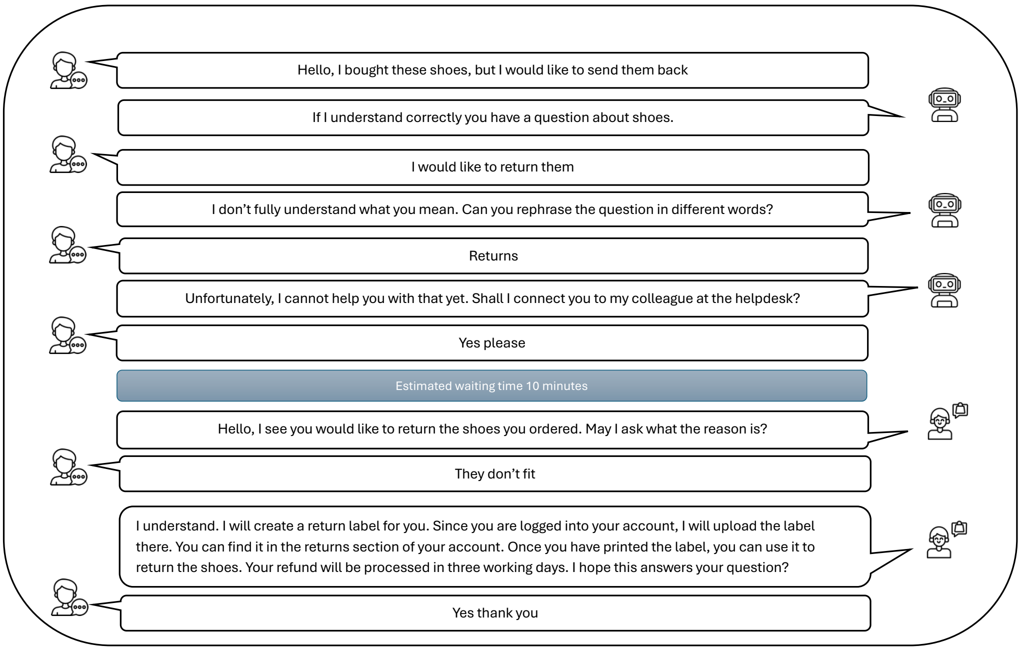
These steps (1-5) will be followed in the discussion of the evaluation of example conversation.

Fluency *(1)* tends to refer to the naturalness of utterances produced by a dialogue system. D'Haro et al. (2019) take the definition of fluency from machine translation and focus on the quality of the construction, emphasising syntax.

(2) Fluency is mostly measured automatically, although incidentally some researchers also ask users about their perceptions of fluency. For example, Firdaus et al. (2020) ask users to rate fluency based on the question 'The generated response is grammatically correct and is free of any errors.' Different automatic metrics can be used to measure fluency. In the literature review we found that, perplexity (used by Firdaus et al. 2020), the BLEU score (used by Peng et al. 2021) and AM-FM (used by D'Haro et al. 2019) were used for automatically measuring fluency. Perplexity refers to the inverse probability of an utterance (computed using a language model), normalised by the number of words in that utterance (Jurafsky and Martin, 2024). Intuitively, this corresponds to the surprisal of seeing a particular sequence of words, given data used to train a language model. Firdaus et al. (2020) emphasise that as the perplexity score decreases, the responses become more fluent and grammatical. BLEU is a n-gram based textual similarity score (Papineni et al., 2002), it needs reference utterances to compare a candidate utterance with (see Section 3.2 for more information about BLEU). BLEU's dependence on context-dependent reference data suggests a notion of 'conversational fluency' which incorporates ideas about appropriateness of the generated response. This is different from what is captured by the perplexity metric, which only looks at the intrinsic fluency of an utterance in isolation, disregarding the context (which is to be captured by a different metric). In addition, BLEU is dependent on its settings (like the amount of references), which complicates the comparison of BLEU scores across studies (Post, 2018). To overcome the limitation of needing a reference text, AM-FM was introduced (Banchs and Li, 2011). The AM-FM metric is an adequacy-fluency metric that uses both syntactic (referencing to fluency) and semantic (referencing to adequacy) information on a sentence level (D'Haro et al., 2019). The fluency component of the metric is measured by using the probability of a n-gram language model.

(3) The example chatbot is made with intent-recognition and pre-scripted responses. Therefore, we would expect fluency mostly to be considered when the responses were designed - if fluency is measured on the utterance level. When fluency is measured with regards to previous and next turns, evaluation should be performed after the conversation has taken place. The pre-scripted responses can not be adapted to certain situations, therefore they may not be completely in line with what the user was asking for.

platform, though at a smaller scale.



Hello, I bought these shoes, but I would like to send them back

If I understand correctly you have a question about shoes.

I would like to return them

I don't fully understand what you mean. Can you rephrase the question in different words?

Returns

Unfortunately, I cannot help you with that yet. Shall I connect you to my colleague at the helpdesk?

Yes please

Estimated waiting time 10 minutes

Hello, I see you would like to return the shoes you ordered. May I ask what the reason is?

They don't fit

I understand. I will create a return label for you. Since you are logged into your account, I will upload the label there. You can find it in the returns section of your account. Once you have printed the label, you can use it to return the shoes. Your refund will be processed in three working days. I hope this answers your question?

Yes thank you

Figure 7: A fictitious example of a customer service interaction. (Icons from Freepik.com.)

Both perplexity and AM-FM use external language models to score fluency. Additionally, they operate at the sentence level. One can therefore wonder how these metrics can be applied to our example as the answers of the chatbot are predefined. In theory, these metrics can be applied in the design phase, however, intuition of the designer might be an equally effective measure for fluency. BLEU on the other hand does consider to some degree the context as the human-generated reference answers are typically context-dependent. However, in the case of the pre-scripted example the answers are already human-generated and therefore can be considered the reference data.

(4 + 5) As became evident from the discussion above, automatically evaluating fluency in this context is probably not meaningful as the responses are pre-defined. As described above, in some cases human evaluation is used to determine the fluency of chatbot utterances. In the case of this example, humans wrote the pre-defined responses. However, a chatbot is not static and continuous development is needed to keep it up-to-date. Therefore, designers need to keep fluency in mind when refining and creating responses.

Effectiveness *(1)* measures the extent to which the system achieves its intended results, also considering the process of achieving that result (which relates closely to efficiency).

(2) In the literature review, time needed to complete the task and the correctness of the task were for example used to assess the dialogue systems effectiveness (Okanović et al., 2020; Tsai et al., 2022). Time needed can also be applied as a measure of efficiency, however the combination with correctness of the final task aligns it closer to effectiveness. Effectiveness is sometimes also measured by employing questionnaires. For this example, only the time needed might not be a sufficient indication of effectiveness as there is also a handover to a human employee. Therefore, obtaining the users' perceptions of effectiveness might be a valuable addition.

(3) In our example, time needed to complete the task is quite long as the chatbot needs to handover to an employee. Eventually the employee is able to solve the user's request. However, the chatbot is not able to do this so we could argue that the chatbot is not very effective. Companies also often measure effectiveness based on these handovers. This is what is often called the escalation rate. [22] The less escalations to a human employee, the more effective the chatbot is. However, this might not in all cases be a good measure of effectiveness. Some users might have left the chatbot early without having their problem solved. These users are not counted as escalations, although the chatbot was not able to handle their request. In these cases the chatbot's effectiveness is still low.

(4 + 5) In the case of our example one can wonder if it is an effective conversation if we base effectiveness on the time needed. Therefore, it might be im-

[22] See for example the KPI in the Microsoft Bot dashboard https://learn.microsoft.com/en-us/dynamics365/customer-service/use/oc-bot-dashboard?tabs=copilotstudiodashboard.

portant to also get the user's perceptions on the topic. In our review we saw that measures like time are sometimes combined with a questionnaire item. Pricilla et al. (2018) for example combines the number of tasks a user was able to complete with a questionnaire item that asks the user to rate how effective the system was to use. In the case of our example this question could be interpreted in multiple ways. Maybe participants find the chatbot effective because it hands over, maybe others would find this not effective at all. Therefore, the results of such a single item might be slightly hard to interpret. This means that multiple items need to be included, the item needs to be specified, or the user needs to provide a explanation with their rating.

Satisfaction *(1)* was also discussed in Section 3.1. With satisfaction one can measure if users are satisfied with the conversation, an utterance and/or the system as a whole.

(2) In our literature review satisfaction was only measured by means of human evaluation (see Section 3.2). However, there have been attempts to automatically predict user satisfaction, for example by making a prediction model based on human-annotated data (e.g. Sun et al. 2021; Lin et al. 2024). Nevertheless, one can wonder if satisfaction can be thoroughly measured automatically from a conversation such as the one in Figure 7. This conversation might not be rich enough to measure a construct like satisfaction (for a discussion of conversational richness and measurement see Van Miltenburg et al. 2025).

(3) In this example the customer remains quite polite, but at the same time the customer might not be very happy with the chatbot's performance. Therefore, we can argue that human evaluation of satisfaction is necessary. Two existing scales (as discussed in Section 3.2) were used in the literature we discussed: the System Usability Scale (SUS) (Brooke, 1996) and the Questionnaire for User Interface Satisfaction (QUIS) (Chin et al., 1988). In this case we should check how well the questionnaires line up with our chatbot and what we intend to measure with satisfaction. If the focus in on the interface we could opt for the QUIS, if the focus is broader (which would be the case in this example) we could opt for SUS. In Section 3.2 an example item of the SUS was already mentioned: 'I thought the system was easy to use'. Some of the items might need to be adapted to the current situation, for example by changing system into chatbot. As mentioned before, this construct is multi-faceted and we would expect multiple items in this case for measuring satisfaction. At the same time, next to only rating these items on a scale we would like to add open-ended questions if possible to elicit why for example the system was (or was not) easy to use. This would give more directions for improving the chatbot.

(4+5) Regarding the example conversation, we could imagine that the user would be satisfied at the end of the conversation. After all, the employee was able to help the customer with the query. But if the focus is purely on the functioning of the chatbot, the customer would not be so satisfied any more – the employee saved this conversation. The handover option is probably preferred over an error. Therefore, it would be good in this case to at least split the satisfaction with the overall conversation from the satisfaction with the chatbot conversation. It is therefore important to phrase the questionnaire items accordingly to emphasise that it concerns satisfaction regarding the chatbot.

Conclusion This example shows that it is important to first determine what constructs need to be measured. For this case study we chose three constructs that were measured differently in literature but might not be the most informative in a certain context. Next, it is important to consider how these constructs should be measured and in what phase of the development. A construct like fluency should be tackled in this case maybe even before deployment and during the iterative development process, while effectiveness can only be measured after the conversation has taken place. Additionally, when conversation logs are used, it is important to consider what can actually be measured from these logs. This example shows that there is not one universal evaluation approach, but that the choice for the constructs and metrics depend on the domain and context of the system. With the steps outlined in this section we hope to provide a starting point for researchers and practitioners for evaluating their dialogue system.

6 Key Challenges of Applying Evaluation in the Customer Service Domain

Different domains all have specific characteristics and thus use (and need) different constructs and subsequently different evaluation metrics. In this section the customer service domain is again used as a case study to show the importance of taking into account the context in which a system is employed. In customer service, chatbots are increasingly employed and constantly under development (both in scientific and practical settings). Costello and LoDolce (2022) predict that by 2027, chatbots will become the primary communication channel for a quarter of organisations. Therefore, there is a need to create an overview of evaluation methods and constructs for this task-oriented domain.

Within customer service, dialogue systems do not just interact with users to answer their questions or help with basic tasks (i.e., task-based systems), but they also serve as *brand ambassadors* (see Harris and de Chernatony 2001 for a discussion of this concept). Both users and organisations perceive customer service chatbots as an extension - and sometimes even as a (partial) replacement - of the human customer service agent with whom (potential) customers could chat or call (Zhang et al., 2023). Thus, whatever these systems do also reflects on the corporate image of the organisation that they serve. Bad experiences with a dialogue system may give (potential) customers a bad impression of the organisation as a whole (Liebrecht et al., 2024).

In addition, customer service chatbots have a set of characteristics that distinguish them from other chatbot contexts like health care or social chatbots. Users look for an efficient conversation (Brandtzaeg and Følstad, 2018). They expect to be assisted in a friendly, often human like, manner (Liebrecht and van der Weegen, 2019; Liebrecht et al., 2021). In all cases, a user (customer) has a certain task that needs to be accomplished with a (task-based) chatbot (Zhang et al., 2020b). Often the conversations are text-based and last several turns until the query of the customer is answered (or if unsuccessful, the conversation results in a breakdown; see Braggaar et al. 2024). When the conversation is finished, the customer has formed an opinion not only about the dialogue system itself but often also about the organisation that the chatbot represents. Pavone et al. (2023) for example show that customers blame the company more for negative outcomes than the chatbot itself. Thus, as good evaluations of organisations are important for their image, negative user experiences of a customer service chatbot conversation should be avoided.

Figure 8 provides a general model of customers interacting with a chatbot that acts on behalf of an organisation (see Fitrianie et al. 2020 for a similar model on interactions between humans and artificial social agents). The customer accesses the chatbot with a particular set of goals that they would like to achieve, and the chatbot is designed to help the customer achieve (a subset of) those goals, as a means to lighten the load on the human customer service agents. Those agents and the chatbot collaborate to provide customers with the best possible service, to create a positive impression of the organisation.[23] We can see the different constructs that have been discussed so far as relating to our general model of chatbot interaction. Specifically: most metrics either focus on the intrinsic qualities of the chatbot, or the user's opinion of the chatbot. This leaves open many research questions about the rest of Figure 8. No papers in our sample looked into the experiences of human support agents, or into the user's opinion about the organisation represented by the chatbot. At the same time, researchers in communication science and human-computer interaction are building up a body of knowledge about the different relations illustrated in Figure 8. We believe it would be very useful to build a connection between NLP and these neighbouring fields. To this end we provide some pointers below.

6.1 The Relation between Users and Chatbots

Now organisations increasingly implement chatbots in their customer service, and customers are at the first instance exposed to this automated interlocutor when seeking for online assistance, it is of great importance that customers accept the technology. Several existing theories describe which factors impact users' acceptance of new technologies.

The Technology Acceptance Model (Davis, 1989, TAM) is a widely used theoretical framework that seeks to explain and predict how users accept and use new technologies. It is based on the premise that the perceived usefulness and perceived ease of use of a technology are key factors influencing its adoption - in this literature review we saw that both constructs were also distinguished as relevant measures for the evaluation of task-oriented dialogue systems.

Over the years, several extensions of TAM were presented to provide a more comprehensive understanding of technology acceptance, such as TAM2 (Venkatesh and Davis, 2000), the Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003, UTAUT), and subsequently UTAUT2 (Venkatesh et al., 2012). The theoretical models distinguish additional predictors of perceived usefulness and ease of use, such as job relevance (the extent to which the user believes the system is suitable for the job; Venkatesh and Davis 2000), output quality (the perception of the system's ability to perform specific tasks; Venkatesh and Davis 2000), and hedonic motivation (user's experience of joy and playfulness; Venkatesh et al. 2012). These predictors can be related to evaluation constructs identified in the current literature review, such as *competence*, *usefulness*, and *enjoyment*.

Theoretical models on technology acceptance thus provide both (NLP)researchers and chatbot developers grip on the constructs to (systematically) focus on when evaluating a task-based dialogue system to gain insight into the relation between user and chatbot.

[23] Of course, at the organisation itself there are also different stakeholders involved in the development, maintenance, and day-to-day operation of the chatbot. We will not discuss these in detail, to avoid further complication.

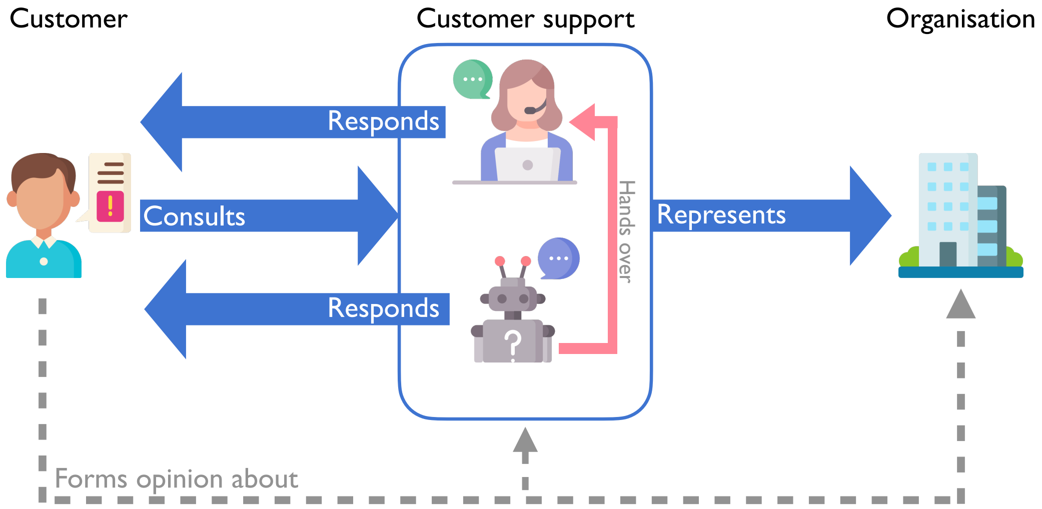


Figure 8: A general model of customers interacting with a chatbot that acts on behalf of an organisation. By interacting with the chatbot, customers form impressions and opinions about both the chatbot and the organisation. Some conversations cannot be handled by the chatbot alone, and should be handed over to a human agent who then responds to the customer. (Icons from Freepik.com.)

6.2 The Relation between Users and Organisations

In the field of communication science, several theoretical models have been developed that describe how people's initial expectations towards e.g., an organisation, a product or service, or a communicative utterance, can impact their final satisfaction and subsequently their continuance intentions.

One of the theoretical models that can describe the relation between user and organisation in the current study's context, is the Expectation-Confirmation Theory (Oliver, 1980, ECT). The ECT originated in consumer behaviour research and describes how expectations and (dis)confirmation of these expectations of e.g., an organisation's service performance can impact users' satisfaction. The theory has been applied to various technology adoption contexts, among which chatbot research. ECT unfolds through four stages (Oliver and DeSarbo, 1988):

1. Applied to the context of chatbot research, the first stage describes users' expectations about the technology prior to using it. With regard to user expectations, for example, it has been known that they can be shaped by several social characteristics of the chatbot such as conscientiousness, personalisation and emotional intelligence (Chaves and Gerosa, 2021). Also in the subsequent stages of the ECT, several constructs that have been identified in the current literature review can be applied.

2. The second stage describes the usage of the technology itself, which is influenced by these expectations. If a notable disparity arises between the actual performance and user expectations, perceived performance adjusts in accordance (either increasing or decreasing) with those expectations.

3. In the third stage, the perceived performance either aligns with or contradicts user expectations. This has been shown by, for example, Khadpe et al. (2020) who on the basis of three studies state that projecting a chatbot's competence can be beneficial to attract new users, but should be corrected quickly during the real chatbot interaction to avoid discardance.

4. Finally, in the fourth stage, user satisfaction is impacted by the interplay of user expectations and perceived approval levels, with satisfaction increasing when user expectations are met. Satisfaction, in turn, could impact other perception measures, such as the user's intention to reuse the chatbot in the future (e.g., Ashfaq et al. 2020), loyalty towards the organisation (Cheng and Jiang, 2020), and purchase intention (Jiang et al., 2022).

The takeaway here is that the different constructs mentioned above are all related through a general theory of human behaviour. Some are related because they can be subsumed under a more general construct, while others are causally related. For example, if someone is satisfied with a chatbot they are more likely to stay loyal to the organisation and use the system again in the future. Our impression is that NLP researchers are mostly focused on measuring performance, and they

1. Define the construct of interest. Try to ground the definition in the literature on the topic (e.g. linguistics or psychology).
2. Motivate your choice of constructs (why are they relevant?) and show how those constructs relate to each other.
3. Think about the operationalisation of the construct. How do your metrics align with the definition of the construct? Which aspects does it capture?
4. Maximise the comparability of your measurements by using established metrics, but remain critical about their operationalisation.
5. Be detailed and specific. Make sure readers (readers of your paper and evaluators) don't need to fill in any details themselves. Be clear in your formulations.
6. If possible, share your materials, such as the used surveys or the code for automatic metrics. For human evaluation, you could for example use the human evaluation data sheet (HEDS) (Shimorina and Belz, 2022).
7. Focus on generalisability. Reflect on how the evaluation generalises to other contexts/situations.
8. Make your evaluation outcomes public and reflect on your outcomes and methods. This enables scholars to compare results and develop validated measures.

Table 4: Recommendations for evaluation of dialogue systems.

spend relatively little time discussing how different performance measurements may be related.[24]

Existing theory may also help us see whether any constructs may be overlooked in the literature. Previous work focusing on customer service interactions already described some quality measures. Lewis and Mitchell (1990) describe five quality measures (based on work by Parasuraman et al. 1988): tangibles , reliability, responsiveness, assurance, and empathy. Four out of five are directly found in this review, only tangibles (which concerns physical attributes and facilities like the employee) is not clearly found. Ghosh and Mandal (2020) distinguish nine different constructs (or dimensions) that are important in a webcare context (a written human-human service context). Some of these such as coherence and assurance are also found in this review, others such as ownership are not found in this selection of literature. Previous work focusing on customer service in general can thus also help define constructs of interest from their more general customer service perspective.

6.3 The Relation between Human and Automatic Customer Service Agents

Next to the interaction between users and dialogue systems, Figure 8 also shows how human customer service agents are impacted by automation: after the dialogue system triages the customer's request, the interaction may be handed over to human agents. This shows that not only users, but multiple different groups of people are involved in the development and usage of the chatbot. Customer service managers, conversational designers and human agents are all involved with the system, and all of these groups have different perspectives when it comes to evaluation (Martijn et al., 2024). Different perspectives thus might need a different focus when it comes to evaluation. For a full understanding of the performance of a dialogue system, organisations should also evaluate for example the customer service agents' impressions of and experiences with the system. Improving handovers between dialogue systems and human agents may also involve summarising the conversation so far, which can either be framed as a separate task (dialogue summarisation, see Feng et al. 2022) or as part of a conversation with the human agent.

7 Limitations of this Review

No review can ever be fully exhaustive, and our review is similarly limited. We aimed for a transparent selection procedure of the papers, so that our process is reproducible and gaps in our review are easier to identify. As a consequence of the scale and nature of this project (manually analysing all selected publications), our main selection of papers is limited to those published up to 2021. We have addressed this gap by discussing newer developments with respect to LLMs. Nevertheless this paper provides an elaborate overview of most, if not all constructs that are currently being considered in the literature. Furthermore, we focus explicitly on a critical analysis of construct definitions and their operationalisations. This method is timeless and can be applied to all constructs, either already used in literature or newly defined. And although the technology is evolving rapidly, the development of theory on dialogue systems evolves at much slower pace. Multiple experimen-

[24] Work on ethics and AI safety may be the exception here; by the nature of this area, one has to consider the impact of technology on both individual humans as well as on society as a whole. See Abercrombie et al. (2023) for a recent example.

1. How do existing constructs relate to each other?
2. Can the set of constructs be reduced to a smaller set? See for example Fitrianie et al. (2020) on creating a list of unifying questionnaire constructs.
3. How reproducible are different evaluation metrics?
4. How can we overcome terminological confusion with regards to the definition and operationalisation of constructs?
5. How do automatic metrics relate to human evaluation? Can we automatically predict human ratings?
6. How do we know that participants' ideas about a construct correspond to our notion of the construct?
7. Different questions are used for the same construct. What effect do different formulations have on the outcomes?
8. Are new metrics/constructs needed with the advent of LLMs?
9. How can we overcome concerns around the validity and reliability of evaluation using LLMs?
10. How do already developed constructs and metrics line up with existing research in a certain domain such as customer service?
11. Can we predict how users/customers react to a dialogue system based on the existing evaluation methods?
12. How do objective measures correspond to "perceived" measures of a construct?
13. What are the current practices and assumptions among dialogue system researchers with respect to evaluation? Similar to Zhou et al. (2022), who explore this for NLG evaluation, an overview can be made for dialogue system evaluation.

Table 5: Outstanding questions for the evaluation of dialogue systems.

tal studies are needed to test the hypotheses about how different constructs relate to each other.

We were also limited by the amount of documentation provided by the authors of the papers in our selection (the lack of clarity in papers is also reported by Howcroft et al. 2020). Where some papers were extremely detailed and all information could be easily extracted, other papers were not. We have tried to obtain all possible information from the papers (such as construct names, definitions, metrics), but this sometimes proved to be a difficult task.

Because of space constraints, we have not been able to discuss all 109 constructs. Our goal with this review was not to be exhaustive (that would have been impossible and this paper would have become unreadable), but to give an outline of a general method to critically analyse the operationalisation of any construct. Nonetheless, an overview of all constructs can be found on OSF.

8 Conclusion and Future Directions

We set out to provide a systematic review of evaluation methods that are used to assess the performance of task-based dialogue systems. Our results show a wide diversity in both constructs that are considered and evaluation methods that are used in the evaluation of dialogue systems. Next to the diversity in approaches, we also found inconsistencies in the terminology used to refer to different constructs, missing definitions, and an overall lack of detail in the description of the evaluation procedures. This made comparing papers a challenging and time consuming task. To be sure: it should *not* be this hard to determine whether two studies look at the same or different constructs. Moreover, it *should* be straightforward to understand and build on evaluation procedures used in previous research. This problem is not unique to NLP; transparency and reproducibility are recurring themes in the Open Science movement.[25] We are hopeful that in the future researchers will work towards improved reporting standards. To this end we have provided some recommendations in Table 4.

In Sections 3.3 and 4.5 we have discussed recent developments concerning the usage of LLMs. These sections reflected on the potentials of using LLMs to evaluate systems with LLMs. In the context of evaluation, we discussed the current state of research and possible problems with validity concerning the usage of LLMs. We argue that, although LLMs give us many new options to build and evaluate task-oriented dialogue systems, the constructs identified in our review remain relevant. The arrival of LLMs has just meant that some constructs have become more relevant (hallucination, repetition), and there are more ways to operationalise a construct of interest.

Looking at task-oriented dialogue systems at a high level (as in Figure 8), it is clear that there is space for more discussion and innovation regarding the more

[25] For further reading on the Open Science movement, see for example Spellman et al. (2018).

practical applications of evaluation. Our review shows that virtually all attention in the NLP literature goes towards the relation between users and chatbots. But at the same time, there has been an increase in the research about dialogue systems in the customer service domain. We have argued that research in NLP could very well align more with this research. As a field we could at least learn more about theory that has already been developed by scholars in marketing, communication science and human-computer interaction. For example, are we able to relate evaluation metrics developed by NLP researchers to variables that are theoretically significant to communication scientists? And are we able to predict the real-world reactions of customers to a deployed dialogue system?

If anything, our review shows that there are still many outstanding questions regarding the evaluation of dialogue systems. Table 5 provides an overview of the outstanding questions arising from this review. We hope that these will be helpful to guide future research towards a more integrated account of task-oriented dialogue system evaluation.

Acknowledgements

The authors would like to thank Linwei He for annotating a set of papers. This paper is part of the NWO-funded Smooth Operators project (KIVI.2019.009).

References

Abd-Alrazaq, Alaa, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, Kerstin Denecke, et al. 2020. Technical metrics used to evaluate health care chatbots: Scoping review. *Journal of Medical Internet Research*, 22(6):(e18301) 1–15.

Abercrombie, Gavin, Amanda Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages. On anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Abu Shawar, Bayan and Eric Atwell. 2007. Different measurement metrics to evaluate a chatbot system. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pages 89–96, Rochester, NY. Association for Computational Linguistics.

Abu Shawar, Bayan and Eric Atwell. 2016. Usefulness, localizability, humanness, and language-benefit: Additional evaluation criteria for natural language dialogue systems. *International Journal of Speech Technology*, 19(2):373–383.

Akkasi, Abbas, Kathleen Fraser, and Majid Komeili. 2023. Reference-free summarization evaluation with large language models. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 193–201, Bali, Indonesia. Association for Computational Linguistics.

Algherairy, Atheer and Moataz Ahmed. 2024. A review of dialogue systems: current trends and future directions. *Neural Computing and Applications*, 36(12):6325–6351.

Alhadreti, Obead and Pam Mayhew. 2018. Rethinking thinking aloud: A comparison of three think-aloud protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA. Association for Computing Machinery.

Ananthakrishnan, R, Pushpak Bhattacharyya, M Sasikumar, and Ritesh M Shah. 2007. Some issues in automatic evaluation of English-Hindi MT: More blues for BLEU. In *Proceedings of the 5th International Conference on Natural Language Processing(ICON-07)*, Hyderabad, India.

Argyle, Lisa P, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Ashfaq, Muhammad, Jiang Yun, Shubin Yu, and Sandra Maria Correia Loureiro. 2020. I, chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics*, 54:(101473) 1–17.

Aust, Harald and Hermann Ney. 1998. Evaluating dialog systems used in the real world. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 1053–1056. IEEE.

Banchs, Rafael E. and Haizhou Li. 2011. AM-FM: A semantic framework for translation quality assessment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 153–158, Portland, Oregon, USA. Association for Computational Linguistics.

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Belz, Anya, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

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, 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.

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.

Bickmore, Timothy and Toni Giorgino. 2006. Health dialog systems for patients and consumers. *Journal of Biomedical Informatics*, 39(5):556–571.

Braggaar, Anouck, Jasmin Verhagen, Gabriëlla Martijn, and Christine Liebrecht. 2024. Conversational repair strategies to cope with errors and breakdowns in customer service chatbot conversations. In *Chatbot Research and Design*, pages 23–41, Cham. Springer Nature Switzerland.

Brandtzaeg, Petter Bae and Asbjørn Følstad. 2018. Chatbots: Changing user needs and motivations. *Interactions*, 25(5):38–43.

Brooke, John. 1996. Sus: A 'quick and dirty' usability scale. *Usability Evaluation in Industry*, 189(3):189–194.

Bryman, Alan. 2012. *Social research methods*, 4 edition. Oxford University Press, London, England.

Caldarini, Guendalina, Sardar Jaf, and Kenneth McGarry. 2022. A literature survey of recent advances in chatbots. *Information*, 13(1):(41) 1–22,.

Campillos-Llanos, Leonardo, Catherine Thomas, Éric Bilinski, Antoine Neuraz, Sophie Rosset, and Pierre Zweigenbaum. 2021. Lessons learned from the usability evaluation of a simulated patient dialogue system. *Journal of Medical Systems*, 45(7):1–20.

Campillos-Llanos, Leonardo, Catherine Thomas, Eric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Casas, Jacky, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2020. Trends & methods in chatbot evaluation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 280–286.

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799:1–75.

Chaves, Ana Paula and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

Chen, Guiming Hardy, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or LLMs as the judge? A study on judgement biases. *CoRR*, abs/2402.10669:1–27.

Chen, Yi-Pei, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024b. Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13650–13665, Torino, Italia. ELRA and ICCL.

Cheng, Yang and Hua Jiang. 2020. How do AI-driven chatbots impact user experience? Examining gratifications, perceived privacy risk, satisfaction, loyalty, and continued use. *Journal of Broadcasting & Electronic Media*, 64(4):592–614.

Chin, John P, Virginia A Diehl, and Kent L Norman. 1988. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 213–218.

Chung, Minjee, Eunju Ko, Heerim Joung, and Sang Jin Kim. 2020. Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, 117:587–595.

Chung, Willy, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. 2023. InstructTODS: Large language models for end-to-end task-oriented dialogue systems. In *Proceedings of the Second Workshop on Natural Language Interfaces*, pages 1–21, Bali, Indonesia. Association for Computational Linguistics.

Colby, Kenneth Mark, James B. Watt, and John P. Gilbert. 1966. A computer method of psychotherapy: Preliminary communication. *The Journal of Nervous and Mental Disease*, 142(2):148–152.

Costello, Katie and Matt LoDolce. 2022. Gartner predicts chatbots will become a primary customer service channel within five years. https://www.gartner.com/en/newsroom/press-releases/2022-07-27-gartner-predicts-chatbots-will-become-a-primary-customer-service-channel-within-five-years. [Accessed June 14, 2023].

Cronbach, L.J. and P.E. Meehl. 1955. Construct validity in psychological tests. *Psychol Bull*, 52(4):281–302.

Cui, Fuwei, Qian Cui, and Yongduan Song. 2020. A survey on learning-based approaches for modeling and classification of human–machine dialog systems. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1418–1432.

Davidson, Sam, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. 2023. User simulation with large language models for evaluating task-oriented dialogue. *arXiv*, 2309.13233.

Davis, Fred D. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

De Wit, Jan. 2024. Leveraging large language models as simulated users for initial, low-cost evaluations of designed conversations. In *Chatbot Research and Design*, pages 77–93, Cham. Springer Nature Switzerland.

DeLone, William H and Ephraim R McLean. 1992. Information systems success: The quest for the dependent variable. *Information Systems Research*, 3(1):60–95.

Deng, Yang, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Deng, Yang, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. 2025. Proactive conversational AI: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3):1–45.

Denzin, Norman K. 1970. *The research act: A theoretical introduction to sociological methods*. Chicago: Aldine.

Deriu, Jan, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1):755–810.

D'Haro, Luis Fernando, Rafael E Banchs, Chiori Hori, and Haizhou Li. 2019. Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. *Computer Speech & Language*, 55:200–215.

Dingemanse, Mark and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5614–5633, Dublin, Ireland. Association for Computational Linguistics.

Dror, Rotem, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Duggenpudi, Suma Reddy, Kusampudi Siva Subrahamanyam Varma, and Radhika Mamidi. 2019. Samvaadhana: A Telugu dialogue system in hospital domain. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 234–242, Hong Kong, China. Association for Computational Linguistics.

Edwards, Jack L and James A Mason. 1988. Evaluating the intelligence in dialogue systems. *International Journal of Man-Machine Studies*, 28(2-3):139–173.

Eren, Berrin Arzu. 2021. Determinants of customer satisfaction in chatbot use: Evidence from a banking application in Turkey. *International Journal of Bank Marketing*, 39(2):294–311.

Eric, Mihail, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Fan, Mingming, Serina Shi, and Khai N Truong. 2020. Practices and challenges of using think-aloud protocols in industry: An international survey. *Journal of Usability Studies*, 15(2).

Fan, Yifan and Xudong Luo. 2020. A survey of dialogue system evaluation. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1202–1209. IEEE.

Federici, Stefano, Maria Laura de Filippis, Maria Laura Mele, Simone Borsci, Marco Bracalenti, Giancarlo Gaudino, Antonello Cocco, Massimo Amendola, and Emilio Simonetti. 2020. Inside pandora's box: A systematic review of the assessment of the perceived quality of chatbots for people with disabilities or special needs. *Disability and Rehabilitation: Assistive Technology*, 15(7):832–837.

Feng, Xiachong, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5453–5460. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Finch, James D., Sarah E. Finch, and Jinho D. Choi. 2021. What went wrong? Explaining overall dialogue quality through utterance-level impacts. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 93–101, Online. Association for Computational Linguistics.

Finch, Sarah E. and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Firdaus, Mauajama, Arunav Pratap Shandeelya, and Asif Ekbal. 2020. More to diverse: Generating diversified responses in a task oriented multimodal dialog system. *PloS one*, 15(11):(e0241271) 1–26.

Fitrianie, Siska, Merijn Bruijnes, Deborah Richards, Andrea Bönsch, and Willem-Paul Brinkman. 2020. The 19 unifying questionnaire constructs of artificial social agents: An IVA community analysis. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, pages 1 – 8, New York, NY, USA. Association for Computing Machinery.

Flake, Jessica Kay and Eiko I. Fried. 2020. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4):456–465.

Følstad, Asbjørn, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: An interdisciplinary research agenda. *Computing*, 103(12):2915–2942.

Fonseca, Erick, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Foster, Mary Ellen, Manuel Giuliani, and Alois Knoll. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 879–887, Suntec, Singapore. Association for Computational Linguistics.

Fried, Eiko I and Jessica K Flake. 2018. Measurement matters. *APS Observer*, 31(3):29–31.

Gehrmann, Sebastian, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghav 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, …, 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.

Gehrmann, Sebastian, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, …, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Ghosh, Tathagata and Santanu Mandal. 2020. Webcare quality: Conceptualisation, scale development and validation. *Journal of Marketing Management*, 36(15-16):1556–1590.

Gonzales, Hada M Sánchez and Marıa Sánchez González. 2017. Bots as a news service and its emotional connection with audiences. The case of politibot. *Doxa Comunicación: Revista interdisciplinar de estudios de Comunicación y Ciencias Sociales*, 25:51–68.

Gururangan, Suchin, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? Measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Han, Lifeng, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Harms, Jan-Gerrit, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2018. Approaches for dialog management in conversational agents. *IEEE Internet Computing*, 23(2):13–22.

Harris, Fiona and Leslie de Chernatony. 2001. Corporate branding and corporate brand performance. *European Journal of Marketing*, 35(3/4):441–456.

Hart, Sandra G and Lowell E Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier.

Holmes, Samuel, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael McTear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 207–214.

Horton, John J. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Howard, George S. 1981. On validity. *Evaluation Review*, 5(4):567–576.

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.

Hu, Xinyu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are LLM-based evaluators confusing NLG quality criteria? *CoRR*, abs/2402.12055.

Huang, Hui, Shuangzhi Wu, Xinnian Liang, Bing Wang, Yanrui Shi, Peihao Wu, Muyun Yang, and Tiejun Zhao. 2023. Towards making the most of LLM for translation quality estimation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 375–386. Springer.

Jannach, Dietmar, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36.

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):1–38.

Jiang, Hua, Yang Cheng, Jeongwon Yang, and Shanbing Gao. 2022. AI-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior. *Computers in Human Behavior*, 134:(107329) 1–14.

Jurafsky, Daniel and James H. Martin. 2024. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models. Retrieved October 29, 2024, from https://web.stanford.edu/ jurafsky/slp3/. Online manuscript released August 20, 2024.

Karpinska, Marzena, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kataoka, Yuki, Tomoyasu Takemura, Munehiko Sasajima, Naoki Katoh, et al. 2021. Development and early feasibility of chatbots for educating patients with lung cancer and their caregivers in Japan: Mixed methods study. *JMIR cancer*, 7(1):(e26911) 1–7.

Khadpe, Pranav, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

Kimchi, Judith, Barbara Polivka, and Joanne Sabol Stevenson. 1991. Triangulation: Operational definitions. *Nursing Research*, 40(6).

Kocaballi, Ahmet Baki. 2023. Conversational AI-powered design: ChatGPT as designer, user, and product. *CoRR*, abs/2302.07406:1–36.

Kocmi, Tom and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Kusal, Sheetal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Sashikala Mishra, and Ajith Abraham. 2022. Ai-based conversational agents: A scoping review from technologies to future directions. *IEEE Access*, 10:92337–92356.

Lazarsfeld, Paul F. 1958. Evidence and inference in social research. *Daedalus*, 87(4):99–130.

Leiter, Christoph, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The Eval4NLP 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.

Leusmann, Jan, Chao Wang, and Sven Mayer. 2024. Comparing rule-based and LLM-based methods to enable active robot assistant conversations. https://cui.acm.org/workshops/CHI2024/index.php/accepted-papers/. Accessed: 05-03-2025.

Lewis, Barbara R and Vincent W Mitchell. 1990. Defining and measuring the quality of customer service. *Marketing Intelligence & Planning*, 8(6):11–17.

Li, Zhen, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for NLG evaluation: A survey. *CoRR*, abs/2401.07103.

Liebrecht, Christine, Emiel van Miltenburg, Charlotte van Hooijdonk, Florian Kunneman, Anouk Merckens, and Nik Niessen. 2024. Hoe halen chatbots de kink uit de kabel? *Tijdschrift voor Communicatiewetenschap*, 52(3):288–325.

Liebrecht, Christine, Lena Sander, and Charlotte van Hooijdonk. 2021. Too informal? How a chatbot's communication style affects brand attitude and quality of interaction. In *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4*, pages 16–31. Springer.

Liebrecht, Christine and Evi van der Weegen. 2019. Menselijke chatbots: Een zegen voor online klantcontact? *Tijdschrift voor Communicatiewetenschap*, 47(3):217–238.

Liesenfeld, Andreas and Mark Dingemanse. 2024. Interactive probes: Action-level evaluation for dialogue systems. *Discourse and Communication.* In press.

Lin, Ying-Chun, Jennifer Neville, Jack Stokes, Longqi Yang, Tara Safavi, Mengting Wan, Scott Counts, Siddharth Suri, Reid Andersen, Xiaofeng Xu, Deepak Gupta, Sujay Kumar Jauhar, Xia Song, Georg Buscher, Saurabh Tiwary, Brent Hecht, and Jaime Teevan. 2024. Interpretable user satisfaction estimation for conversational systems with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11115, Bangkok, Thailand. Association for Computational Linguistics.

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.

Liu, Hong, Yucheng Cai, Zhijian Ou, Yi Huang, and Junlan Feng. 2022. A generative user simulator with GPT-based architecture and goal state tracking for reinforced multi-domain dialog systems. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 85–97, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.

Liu, Zixi, Yang Feng, and Zhenyu Chen. 2021. DialTest: Automated testing for recurrent-neural-network-driven dialogue systems. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 115–126.

Maier, Elisabeth, Marion Mast, and Susann LuperFoy. 1997. Overview. In *Dialogue Processing in Spoken Language Systems*, pages 1–13, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mariani, Marcello M., Novin Hashemi, and Jochen Wirtz. 2023. Artificial intelligence empowered conversational agents: A systematic literature review and research agenda. *Journal of Business Research*, 161:(113838) 1–23.

Maroengsit, Wari, Thanarath Piyakulpinyo, Korawat Phonyiam, Suporn Pongnumkul, Pimwadee Chaovalit, and Thanaruk Theeramunkong. 2019. A survey on evaluation methods for chatbots. In *Proceedings of the 2019 7th International conference on information and education technology*, pages 111–119.

Martijn, Gabriëlla, Charlotte van Hooijdonk, Florian Kunneman, and Hans Hoeken. 2024. Reconfiguring the customer service domain: Perspectives of managers, conversational designers, and human agents on human–chatbot collaboration. *International Journal of Innovation and Technology Management*, 21(04):(2450028) 1–28.

McTear, Michael and Marina Ashurkina. 2024. A new era in conversational AI. In *Transforming Conversational AI: Exploring the Power of Large Language Models in Interactive Conversational Agents*, pages 1–16. Apress, Berkeley, CA.

Messick, Samuel. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749. Place: US Publisher: American Psychological Association.

Meyer, Selina, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E. Losada. 2022. Do we still need human assessors? Prompt-based GPT-3 user simulation in conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA. Association for Computing Machinery.

Miraj, F, H Raza, OA Hussain, DA Siddiqi, A Habib, AJ Khan, and S Chandir. 2021. Development and feasibility-testing of an artificially intelligent chatbot to answer immunization-related queries of caregivers in pakistan: A mixed-methods evaluation. In *Tropical Medicine & International Health*, volume 26, pages 140–141. Wiley 111 River St, Hoboken 07030-5774, NJ USA.

Mitchell, Melanie. 2021. Why AI is harder than we think. *CoRR*, abs/2104.12871:1–12.

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.

Motger, Quim, Xavier Franch, and Jordi Marco. 2022. Software-based dialogue systems: Survey, taxonomy, and challenges. *ACM Comput. Surv.*, 55(5).

Nakano, Mikio, Hironori Takeuchi, Sadahiro Yoshikawa, Yoichi Matsuyama, and Kazunori Komatani. 2025. Dialogue systems engineering: A survey and future directions. *arXiv*, 2508.02279.

Ni, Jinjie, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review*, 56(4):3055–3155.

Nielsen, Janni, Torkil Clemmensen, and Carsten Yssing. 2002. Getting access to what goes on in people's heads? Reflections on the think-aloud technique. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, NordiCHI '02, page 101–110, New York, NY, USA. Association for Computing Machinery.

Noble, Helen and Roberta Heale. 2019. Triangulation in research, with examples. *Evidence-Based Nursing*, 22(3):67–68.

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.

Okanović, Dušan, Samuel Beck, Lasse Merz, Christoph Zorn, Leonel Merino, André van Hoorn, and Fabian Beck. 2020. Can a chatbot support software engineers with load testing? Approach and experiences. In *Proceedings of the ACM/SPEC International Conference on Performance Engineering*, pages 120–129.

Oliver, Richard L. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4):460–469.

Oliver, Richard L. and Wayne S. DeSarbo. 1988. Response determinants in satisfaction judgments. *Journal of Consumer Research*, 14(4):495–507.

Ouzzani, Mourad, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1):1–10.

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.

Parasuraman, A., Valarie A Zeithaml, and Leonard L Berry. 1988. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1):12–37.

Park, Dong-Min, Seong-Soo Jeong, and Yeong-Seok Seo. 2022. Systematic review on chatbot techniques and applications. *Journal of Information Processing Systems*, 18(1):26–47.

Pavone, Giulia, Lars Meyer-Waarden, and Andreas Munzel. 2023. Rage against the machine: Experimental insights into customers' negative emotional responses, attributions of responsibility, and coping strategies in artificial intelligence–based service failures. *Journal of Interactive Marketing*, 58(1):52–71.

Peng, Baolin, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2021. RADDLE: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. 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 4418–4429, Online. Association for Computational Linguistics.

Peng, Zhenhui and Xiaojuan Ma. 2019. A survey on construction and enhancement methods in service chatbots design. *CCF Transactions on Pervasive Computing and Interaction*, 1(3):204–223.

Piau, Antoine, Rachel Crissey, Delphine Brechemier, Laurent Balardy, and Fati Nourhashemi. 2019. A smartphone chatbot application to optimize monitoring of older patients with cancer. *International Journal of Medical Informatics*, 128:18–23.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Pradhan, Abhishek and Ketan Todi. 2023. Understanding large language model based metrics for text summarization. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 149–155, Bali, Indonesia. Association for Computational Linguistics.

Press, Gil. 2023. One negative chatbot experience drives away 30% of customers. https://www.forbes.com/sites/gilpress/2023/02/01/one-negative-chatbot-experience-drives-away-30-of-customers/. Accessed: 2025-02-13.

Pricilla, Catherine, Dessi Puji Lestari, and Dody Dharma. 2018. Designing interaction for chatbot-based conversational commerce with user-centered design. In *2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA)*, pages 244–249. IEEE.

Przegalinska, Aleksandra, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons*, 62(6):785–797.

Rei, Ricardo, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up COMETKIWI: Unbabel-IST 2023 submission for the quality estimation shared task. *arXiv*, 2309.11925.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Reiter, Ehud. 2017. How to do an NLG evaluation: Metrics. https://ehudreiter.com/2017/05/03/metrics-nlg-evaluation/. Published on Ehud Reiter's personal blog. Last accessed 24 October 2023.

Reiter, Ehud. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Reiter, Ehud. 2022. We need to understand what users want. https://ehudreiter.com/2022/08/08/we-need-to-understand-what-users-want/. Published on Ehud Reiter's personal blog. Last accessed 24 October 2023.

Ren, Ranci, John W Castro, Silvia T Acuña, and Juan de Lara. 2019. Evaluation techniques for chatbot usability: A systematic mapping study. *International*

Journal of Software Engineering and Knowledge Engineering, 29(11n12):1673–1702.

Resnik, Philip and Jimmy Lin. 2010. *Evaluation of NLP Systems*, chapter *11*. John Wiley & Sons, Ltd.

Riyadh, Md and M Omair Shafiq. 2023. Towards automatic evaluation of NLG tasks using conversational large language models. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 425–437. Springer.

Roque, Geicianfran da Silva Lima, Rafael Roque de Souza, José William Araújo do Nascimento, Amadeu Sá de Campos Filho, Sérgio Ricardo de Melo Queiroz, and Isabel Cristina Ramos Vieira Santos. 2021. Content validation and usability of a chatbot of guidelines for wound dressing. *International Journal of Medical Informatics*, 151:(104473) 1–6.

Schiavone, Sarah R, Kimberly A Quinn, and Simine Vazire. 2023. A consensus-based tool for evaluating threats to the validity of empirical research. *PsyArXiv*, pages 1–28.

Schlangen, David. 2021. Targeting the benchmark: On methodology in current natural language processing research. 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 670–674, Online. Association for Computational Linguistics.

Schmidtova, Patricia, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Sedoc, João, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2018. ChatEval: A tool for the systematic evaluation of chatbots. In *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*, pages 42–44, Tilburg, the Netherlands. Association for Computational Linguistics.

Sedoc, João, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. ChatEval: A tool for chatbot evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 60–65, Minneapolis, Minnesota. Association for Computational Linguistics.

Sekulić, Ivan, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, André Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based user simulator for task-oriented dialogue systems. *arXiv*, 2402.13374.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sensuse, Dana Indra, Venniesa Dhevanty, Ernestina Rahmanasari, Desy Permatasari, Bimo Eka Putra, Jonathan Sofian Lusa, Muhammad Misbah, and Pudy Prima. 2019. Chatbot evaluation as knowledge application: A case study of PT ABC. In *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–6. IEEE.

Shackel, Brian. 2009. Usability–context, framework, definition, design and evaluation. *Interacting with Computers*, 21(5-6):339–346.

Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin.

Shi, Weiyan, Yu Li, Saurav Sahay, and Zhou Yu. 2021. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3478–3492, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shimorina, Anastasia and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Singh, Sonali Uttam and Akbar Siami Namin. 2025. A survey on chatbots and large language models: Testing and evaluation techniques. *Natural Language Processing Journal*, 10:100128.

Specia, Lucia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. *Quality Estimation for Machine Translation*. Synthesis Lectures on Human Language Technologies. Springer Cham.

Spellman, Barbara A, Elizabeth A Gilbert, and Katherine S Corker. 2018. Open science. *Stevens' handbook of experimental psychology and cognitive neuroscience*, 5:1–47.

Strauss, Milton E and Gregory T Smith. 2009. Construct validity: Advances in theory and methodology. *Annu Rev Clin Psychol*, 5:1–25.

Su, Hui, Xiaoyu Shen, Zhou Xiao, Zheng Zhang, Ernie Chang, Cheng Zhang, Cheng Niu, and Jie Zhou. 2020. MovieChats: Chat like humans in a closed domain. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6605–6619, Online. Association for Computational Linguistics.

Subramonian, Arjun, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett. 2023. It takes two to tango: Navigating conceptualizations of NLP tasks and measurements of performance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3234–3279, Toronto, Canada. Association for Computational Linguistics.

Sugawara, Saku, Pontus Stenetorp, and Akiko Aizawa. 2021. Benchmarking machine reading comprehension: A psychological perspective. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1592–1612, Online. Association for Computational Linguistics.

Sugawara, Saku and Shun Tsugita. 2023. On degrees of freedom in defining and testing natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13625–13649, Toronto, Canada. Association for Computational Linguistics.

Sulem, Elior, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Sun, Kaiser, Adina Williams, and Dieuwke Hupkes. 2023. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. *arXiv*, 2310.17514.

Sun, Weiwei, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2499–2506.

Syvänen, Salla and Chiara Valentini. 2020. Conversational agents in online organization–stakeholder interactions: A state-of-the-art analysis and implications for further research. *Journal of Communication Management*, 24(4):339–362.

Takanobu, Ryuichi, Qi Zhu, Jinchao Li, Baolin Peng, Jianfeng Gao, and Minlie Huang. 2020. Is your goal-oriented dialog model performing really well? Empirical analysis of system-wise evaluation. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 297–310, 1st virtual meeting. Association for Computational Linguistics.

Thompson, David A, Paul R Yarnold, Diana R Williams, and Stephen L Adams. 1996. Effects of actual waiting time, perceived waiting time, information delivery, and expressive quality on patient satisfaction in the emergency department. *Annals of Emergency Medicine*, 28(6):657–665.

Thurmond, Veronica A. 2001. The point of triangulation. *Journal of Nursing Scholarship*, 33(3):253–258.

Treadwell, Donald. 2017. *Introducing communication research: Paths of inquiry*. SAGE.

Tsai, Meng-Han, Cheng-Hsuan Yang, Chen-Hsuan Wang, I Yang, Shih-Chung Kang, et al. 2022. Sema: A site equipment management assistant for construction management. *KSCE Journal of Civil Engineering*, 26(3):1144–1162.

Valizadeh, Mina and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Van Deemter, Kees. 2024. The pitfalls of defining hallucination. *Computational Linguistics*, 50(2):807–816.

Van der Goot, Margot J., Laura Hafkamp, and Zoë Dankfort. 2021. Customer service chatbots: A qualitative interview study into the communication journey of customers. In *Chatbot Research and Design*, pages 190–204, Cham. Springer International Publishing.

Van der Lee, Chris, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:(101151) 1–24.

Van der Wal, Oskar, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2024. Undesirable biases in NLP: Addressing challenges of measurement. *Journal of Artificial Intelligence Research*, 79:1–40.

Van Miltenburg, Emiel, Anouck Braggaar, Emmelyn Croes, Florian Kunneman, Christine Liebrecht, and Gabriella Martijn. 2025. Measure only what is measurable: towards conversation requirements for evaluating task-oriented dialogue systems. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 231–238, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Van Miltenburg, Emiel, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021a. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Van Miltenburg, Emiel, Chris Van der Lee, and Emiel Krahmer. 2021b. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Vazire, Simine, Sarah R. Schiavone, and Julia G. Bottesini. 2022. Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, 31(2):162–168.

Venkatesh, Viswanath and Fred D. Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2):186–204.

Venkatesh, Viswanath, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425–478.

Venkatesh, Viswanath, James Y. L. Thong, and Xin Xu. 2012. Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1):157–178.

Walker, Marilyn A., Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.

Wang, Li, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7(41):1–9.

Wang, Peiyi, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *CoRR*, abs/2305.17926.

Weizenbaum, Joseph. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Weizenbaum, Joseph. 1976. *Computer power and human reason*. W.H. Freeman and Company, New York, NY.

Xiao, Ziang, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. *arXiv*, 2305.14889.

Yao, Lucy and Rian Kabir. 2023. Person-centered therapy (Rogerian therapy) [updated 2023 feb 9]. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-.

Ye, Zheng, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards quantifiable dialogue coherence evaluation. 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 2718–2729, Online. Association for Computational Linguistics.

Yeh, Yi-Ting, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33, Online. Association for Computational Linguistics.

Yi, Zihao, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in LLM-based multi-turn dialogue systems. *arXiv*, 2402.18013.

Yin, Zi, Keng-hao Chang, and Ruofei Zhang. 2017. Deepprobe: Information directed sequence understanding and chatbot design via recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2131–2139.

Zhang, Juliana JY, Asbjørn Følstad, and Cato A Bjørkli. 2023. Organizational factors affecting successful implementation of chatbots for customer service. *Journal of Internet Commerce*, 22(1):122–156.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*, pages 1–43.

Zhang, Zheng, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020b. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena.

Zhou, Kaitlyn, Su Lin Blodgett, Adam Trischler, Hal Daumé III, Kaheer Suleman, and Alexandra Olteanu. 2022. Deconstructing NLG evaluation: Evaluation practices, assumptions, and their implications. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–324, Seattle, United States. Association for Computational Linguistics.