Off the Hamster Wheel: Rethinking Dialogue Research through a Meta-Analysis of the ACL Anthology 2024

Amandine Decker[1,2], **Maxime Amblard**[1], **Ellen Breitholtz**[2]
[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2]University of Gothenburg, CLASP
{amandine.decker, maxime.amblard}@loria.fr
ellen.breitholtz@ling.gu.se

Abstract

In this paper, we take a meta-review approach to investigate how conversation is currently studied in the field by analysing papers from the ACL Anthology 2024. We retrieved 407 papers, which represents about 6.1% of the papers published in the selected venues, and manually reviewed them to determine the conversational task addressed, the corpora used, and the evaluation methods employed. Our analysis leads to several observations. First, dialogue systems represent about half of the papers of the ACL Anthology 2024 while more formal and analytical approaches cover only 12%. Second, many papers provide lacking corpus descriptions, which shows a detachment from the data which becomes a simple tool instead of one of the pillars NLP/CL applications should be based on. Third, the evaluation methods, in particular when it comes to dialogue systems, often do not assess the interactional aspects of these systems or rely on assumptions not backed up from evidence of the dialogue research community. We argue that the field would benefit from a renewed focus on analysis and formal representation of conversation, a richer evaluation culture that includes interactional quality, and more systematic practices regarding the data presentation in papers.

Keywords: Dialogue, Dialogue Systems, Science of Science

1. Introduction

Conversation stands as the primary mode of human communication, facilitating the exchange of ideas, emotions, and intentions. It has been the subject of extensive study across various disciplines, including linguistics, psychology, and computer science, yet it remains a complex and elusive phenomenon.

Meanwhile, the use of new modalities of communication such as messaging and video conferencing tools is increasing (Joskowicz, 2023; Omipidan, 2024), and an increasing amount of conversational AI systems are developed. Additionally, the rise of Large Language Models (LLMs) in the past few years has lead to a surge in dialogue system (DS) development and applications. These new paradigms raise questions about how conversation is conceptualised, represented, and evaluated within the Natural Language Processing (NLP) / Computational Linguistic (CL) community.

Several surveys have been conducted in recent years on different aspects of DSs. Kusal et al. (2022) reviewed the various approaches to AI-based conversational agents, while Castillo-López et al. (2025) examined the turn-taking abilities of spoken DSs. Kwan et al. (2023) examined dialogue policy – the component that determines the system's next action in task-oriented dialogue – and emphasised the need for evaluation metrics that better capture how effectively policies support users in completing their tasks. Similarly, Singh and Namin (2025) provided an overview of chatbot architectures and their associated evaluation methods. Taking a broader perspective on evaluation, Yang et al. (2026) conducted a critical analysis of trends in natural language generation (NLG) evaluation between 2020 and 2025. Using automatic methods to retrieve information from papers published in four major computational linguistics venues, they analysed the tasks addressed and the evaluation approaches employed – including automatic metrics, LLM-as-a-Judge methods, and human evaluation. These studies identified persistent gaps in the field, including the lack of harmonised evaluation processes – an issue that remains relevant today, as discussed in this paper.

Focusing on personalised dialogue generation, Chen et al. (2024) investigated approaches that aim to infer or model the user's persona while ensuring that the agent maintains a coherent and relevant persona throughout the interaction. The authors highlight the difficulty of collecting high-quality datasets for this task, as explicitly displaying a persona is often artificial. They also note that many approaches model the agent's and the user's personas separately, thereby overlooking the inherently interactional nature of dialogue. Finally, they also raise concerns about current evaluation practices, which remain relatively simplistic given the complexity of dialogue generation.

While these surveys provide valuable insights into specific tasks or system components, there remains a need for a more holistic overview of the current state of conversation research within our

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2638–2652
11–16 May 2026.

community and of the role of dialogue generation within it. Indeed, a theoretical understanding of conversation is essential for explaining and predicting how humans interact – particularly in a context marked by misinformation (Lazer et al., 2018; Vijayan et al., 2025) and populism (Algan et al., 2017; Norris and Inglehart, 2019; Kaltwasser and Taggart, 2025), where identifying manipulative or deceptive strategies is critical. Moreover, the design and evaluation of DSs require a formal characterisation of dialogue in order to determine which features should be modelled, select appropriate training data, and conduct meaningful evaluation.

In this paper, we take a meta-review approach to investigate how conversation is currently studied in the field by analysing papers from the ACL Anthology 2024. We focus on papers written in English from major venues indexed in the anthology (excluding workshops) using metadata accessed via the ACL API[1]. Our goal is not to provide an exhaustive survey, but to provide a snapshot of where dialogue research stands today, what tasks are being studied, what resources they rely on, and how these tasks are evaluated. This enables us to identify the current gaps in the field and define future research directions.

We begin by outlining the notion of dialogue (section 2) and our approach to examining the current state of dialogue research (section 3). We then detail the main dialogue-related tasks addressed in the NLP/CL community (section 4), discuss the dialogue resources used and created (section 5), and review widely used evaluation methods for DSs (section 6). Finally, we examine recent research practices and the impact of LLMs on the field (section 7), before concluding with our own recommendations to ensure the quality, relevance, and rigour of dialogue research in the LLM era (section 8).

2. The Nature of Dialogue

Face-to-face conversation is at the core of human communication. It is the first form of interaction we learn as children and remains fundamental to how we connect with one another (Clark, 1996). However, despite its central role, it is difficult to study systematically. Capturing face-to-face conversation in ecologically valid settings presents substantial challenges: conversations are multimodal by nature – encompassing not only language, but also gesture, gaze, prosody, etc. (Mondada, 2019; Kendrick et al., 2023). These extra-linguistic modalities are complicated to record without disturbing the participants, making the conversation less natural. Guaranteeing privacy is also hard in natural settings, due to voice and/or face recording but also as daily life topics of conversation involve sharing personal information which can be used for de-anonymisation purposes (Amblard et al., 2014).

A prototypical conversation includes two to four participants, as more than four typically leads to the emergence of sub-groups of conversation (Dunbar et al., 1995; Dezecache and Dunbar, 2012; Krems et al., 2016; Krems and Wilkes, 2019) and makes dialogue more complex (Traum, 2003). Participants should be within arm's reach of each other, enabling them to see and hear one another, as well as share the same physical environment. This proximity enables them to incrementally interpret the current speaker's utterance, allowing for timely feedback (Schegloff, 1982; Clark and Schaefer, 1992) and the planning of responses. The organisation of turns – turn-taking – is generally smooth, characterised by brief pauses between speech turns and occasional short overlaps (Sacks et al., 1978). Participants should be able to ask clarification questions and correct themselves and each other (practices known as repairs) to maintain the flow and coherence of the dialogue (Jefferson, 1972; Schegloff, 1979; Dingemanse et al., 2015; van Arkel et al., 2020). These interactional mechanisms collectively support the understanding and production processes of the participants.

Due to the complex and multi-faceted nature of natural interaction, research into human-human dialogue should primarily be based on naturalistic corpora. Ideally, these corpora would capture the full richness of face-to-face communication: speech, gaze, gesture, prosody, and shared context. In practice, among others for the reasons stated above, most available datasets consist of degraded or restricted forms of data. For example, we often rely on transcriptions instead of multimodal corpora, task-oriented conversations instead of open-domain ones; we focus on duologues and neglect multilogues, and we build synthetic dialogues reflecting our perception of what a dialogue is rather than its actual properties. However, as various types of restrictions of the data may present problems for particular tasks but not for others, this raises important methodological questions regarding which corpora are suitable for which tasks. Indeed, a conversation produced in a restricted setting compared to face-to-face conversation, may display structural differences. For example, Lücking et al. (2025) show that conversations held between two speakers through a virtual reality setting do not replicate findings on turn-taking and gaze patterns in face-to-face conversation. Such examples indicate that the differences between prototypical face-to-face conversation and a given corpus should be assessed before we can generalise the findings of one experiment to all types of conversation. For this reason, investigating the various

[1] https://aclanthology.org/info/development/

corpora used in the NLP/CL community and the associated tasks could help us identify the current limitations of the data.

3. Method

The ACL ANTHOLOGY[2] is an open-access digital repository of research in CL and NLP. Maintained by the Association for Computational Linguistics (ACL), it hosts conference and workshop proceedings, journal articles, and other scholarly materials dating back to the 1960s and is updated every year. Its open-access nature and structured bibliographic metadata have made it a widely used corpus for bibliometric studies, meta-analyses, and research on scientific trends (Vogel and Jurafsky, 2012; Omodei et al., 2014; Gábor et al., 2016; Pramanick et al., 2023).

In this study, we analyse a targeted subset of the ACL ANTHOLOGY to understand how conversation has recently been studied in the NLP/CL community. We restrict our survey to papers published in 2024 which are indexed under the anthology's major venues. Our selection is presented in table 1. We used the ACL Anthology API to retrieve metadata for all papers published in 2024, excluding venues labelled as workshops in the metadata.

Since our investigation focuses on conversation, we first needed to identify a relevant subset of papers. Because the available metadata does not include keywords, we relied on an alternative filtering strategy based on titles. Specifically, we selected papers whose titles contained the stems "conversation", "dialogue", and "discourse". Although this approach may miss some relevant papers and include a few marginal ones, it provides a practical and transparent method for assembling a representative sample.

An alternative would have been to analyse the abstracts of all papers in the ACL ANTHOLOGY 2024. However, the larger number of publications in 2024 – 6685 in total for our sub-selection of venues – makes manual assessment of their topics unrealistic. While automated keyword searches in abstracts were also possible, the number of papers retrieved through our title-based filtering was already sufficiently large to provide a meaningful snapshot of dialogue research in the field.

After excluding papers that did not actually focus on dialogue – for instance those addressing monological discourse or containing unrelated words such as *conversion* – we obtained a dataset of 407 papers, representing approximately 6.1% of the publications from 2024 in the selected venues. Each paper was then manually reviewed by a single annotator to identify the conversational task addressed, the corpora used, and the evaluation methods employed.

[2] https://aclanthology.org/

Venue	#	%
TACL	3	3.2
EMNLP	52	3.6
ACL	36	3.8
NAACL	29	4.4
EACL	14	5.0
JEP/TALN/RECITAL	8	6.0
LREC/COLING	105	6.8
Findings	112	7.8
INLG	5	9.3
CODI	3	17.6
SIGDIAL	40	60.6
Total	407	6.1

Table 1: Number (#) of retrieved paper for the selected venues and portion (%) out of the entire venue in 2024

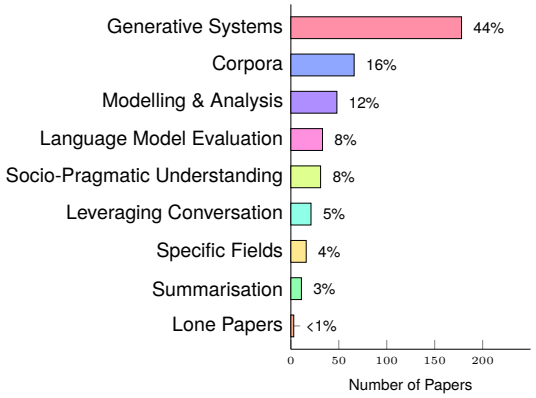

Figure 1: Distribution of papers by task category.

4. Tasks

The NLP/CL field covers a wide range of tasks, so as a first step we annotated each paper with its primary task and grouped these into broader categories. These categories are described below, and their distribution – in both absolute number of papers and as a proportion of the total – is shown in fig. 1. Some papers span multiple categories; for instance, a benchmark proposal may also introduce a new corpus. In such cases, we chose the most salient aspect when assigning categories, so absolute counts should be interpreted cautiously. Nevertheless, the overall distribution captures the dominant research trends in the surveyed set.

Generative Systems Across the ACL ANTHOLOGY 2024, most papers on conversation focus on generative systems (44%). This category encompasses DSs, question answering (QA), and

dialogue state tracking (DST), as well as some works introducing new or improved language models (LMs).

Dialogue systems are programs that interact with humans through natural language. Within this category, TOD systems – e.g. for restaurant booking – are particularly prominent. They assist users in achieving specific goals, with performance often measured by joint goal accuracy, indicating whether all required slots (place, time, etc.) are correctly identified. Less constrained tasks, such as emotional support, negotiation, persuasion, or tutoring, are also explored. For both domain-specific and open-ended dialogues, the conversational dimension remains central but is harder to evaluate (see section 6) than task-specific outcomes.

QA is a subcategory of DSs where users ask questions and receive answers in natural language. Systems can prompt users to refine their queries when answers lack precision, offering more flexibility than web interfaces (Biancofiore et al., 2024). This research area notably focuses on query rewriting and knowledge retrieval.

DST involves monitoring user queries and goals throughout a conversation (Jacqmin et al., 2022). It is pivotal for TOD systems, with recent studies improving DST through zero-shot learning to reduce dialogue collection and annotation costs.

Language Model Evaluation The growing prominence of generative systems in research and society demands robust evaluation to assess their strengths and limitations. Slightly over 8% of the surveyed papers addressed this issue, through benchmark creation or empirical evaluation. Most of these examined LM capabilities across diverse dimensions, from specific skills such as negotiation or slot-value generation to broader conversational dynamics like multi-turn or multi-party dialogue. Security is another concern, with studies probing LMs' risks of leaking private data, susceptibility to misinformation, and tendencies to hallucinate or produce harmful content. Other works explored the limitations of current LMs in languages beyond English.

In some cases, LLMs are used as evaluators, even for tasks involving LLMs themselves. While this enables scalable, context-aware assessment, it also introduces potential biases and limitations in model-based evaluation. We discuss this aspect in more details in section 6.

Socio-Pragmatic Understanding Similarly, the rise of LLMs has heightened the need to identify users' emotional states, communicative intentions, and stances in conversation, as well as subtler aspects such as humour, sarcasm, personality, harmful behaviour, and social meaning recognition. These elements support more contextually aware and socially appropriate DSs. Such tasks, grouped under conversational understanding, accounted for about 8% of the surveyed papers. The SemEval 2024 venue, which we reviewed but excluded for broader generality, also hosted two shared tasks (3 and 10) on conversational emotion recognition.

Summarisation Summarisation is a classical NLP task which remains more complex for conversational- than for non-interactive discourse (Feng et al., 2022). Less than 3% of the papers tackled addressed this task. Most approaches rely on LLMs, with some specifically evaluating hallucinations in generated summaries. Summarisation is particularly prone to such errors (Ramprasad et al., 2024) – as models compress and rephrase content, they often produce plausible but unsupported details.

Leveraging Conversation The recent hype around LLMs has renewed interest in conversational AI, inspiring new directions for traditional tasks, exploiting interactive dialogue to leverage dialogue to create more context-aware, personalised, and adaptive experiences. This category represents slightly over 5% of the surveyed papers. Most reflect emerging paradigms such as conversational recommender systems, where dialogue iteratively refines user queries or suggestions, improving efficiency and user experience.

A key limitation of many such works is their insufficient assessment of dialogue quality, often relying on metrics misaligned with established principles from conversation theory. For instance, the Dist-n metric (Li et al., 2016) measures lexical diversity but offers little insight into coherence, relevance, or contextual appropriateness.

Beyond these common applications, a few studies explore alternative uses of conversation, such as solving complex tasks through multi-turn dialogue (Wang et al., 2024) or generating explanations of news claims (Hsu et al., 2024). Although these papers form a minority, their overlap with DS research may underestimate their actual prevalence in our classification.

Corpus Creation and Resources About 16% of the surveyed papers describe resources or propose methods for resource generation. The datasets range from human–human to human–system dialogues, as well as fully synthetic corpora generated by LLMs and handcrafted ones without real interaction. Several papers present translated or replicated versions of well-known datasets in other languages – reflecting the strong English bias in NLP/CL research and the need for multilingual resources. Other contributions extend or merge existing datasets by adding new annotations, either

human-produced or automatically generated by LLMs. The role of resources in the ACL ANTHOLOGY 2024 and dataset typology are further discussed in section 5.

Modelling and Analysis Approximately 12% of the surveyed papers focus on modelling and analysing conversations, aiming to uncover the mechanisms of human communication rather than to merely build systems. This research advances our understanding of interaction – from turn-taking, feedback, and grounding to topic shifts, redirection strategies, and the conveyance of intentions and information. A well-established task here is dialogue discourse parsing, which segments a conversation into discourse units and identifies relations between them (e.g. question–answer, elaboration, clarification) to model its hierarchical and functional structure (Li et al., 2022). Such insights are crucial for detecting manipulative or deceptive practices, a pressing issue amid widespread misinformation and democratic fragility. They should form the conceptual backbone of conversational research, providing the empirical foundations on which robust and socially responsible systems can be built.

Despite its importance, this line of work remains under-represented at top-tier venues, accounting for only 4% and 6% of papers at EMNLP and ACL (A* in the ICORE Conference Portal), and 3% and 7% at NAACL and EACL (A). Conversely, SigDial, LREC (ranked B[3]) and Findings[4] from ACL, EACL, NAACL, and EMNLP include 10–18% of modelling and analysis papers. This disparity likely stems from the time-intensive nature of such research – requiring detailed annotation, fine-grained analysis, and interdisciplinary collaboration – making it less appealing than system-building.

Specific Fields All the paper categories discussed so far also appear in domain-specific contexts, representing nearly 4% of the surveyed papers, where systems are adapted to particular fields such as medicine and education. In these settings, generative and summarisation models are developed with a targeted focus, often facing additional challenges such as the need for domain-specific data. Several studies address corpus creation for specialised domains (Wang et al., 2024; Arana et al., 2024; Bychkova et al., 2024). While medicine and education are common NLP application areas, each with dedicated workshops indexed in the ACL ANTHOLOGY, one paper even analysed dialogues in motor sports (Isaka et al., 2024), illustrating the field's broad applicability. This diversity highlights the importance of multidisciplinary collaboration, as different domains bring unique knowledge demands, evaluation criteria, and contextual challenges that must be addressed to design effective systems.

Lone Papers The remaining papers (under 1%) present tools for dialogue collection and annotation (Li et al., 2024) or address automatic speech recognition (Lee et al., 2024; Htun et al., 2024).

Because of our venue selection, papers reporting results from shared tasks are not included in this survey. However, such tasks play a crucial role in the community, driving progress across diverse areas, including those under-represented in the venues we covered.

5. Datasets and Resources

Our review considered two corpus-related aspects: the datasets *created* and those *used*.

5.1. Created Datasets

Regarding dataset creation, about one fifth of the surveyed papers introduced a new dataset. Over half are in English, though other languages are represented. In more than a third of papers describing new resources, the language is not explicitly stated but must be inferred from examples or from collection and evaluation methods, and some mention it only in appendices or limitations, despite its importance. English dominates overall, but as repeatedly noted (Bender, 2011; Ducel et al., 2022), it cannot stand for all languages, and the study language should always be explicitly stated.

Some papers address multiple languages simultaneously, covering up to 150 in the XSGD dataset (Tu et al., 2024) – a translation of the English Schema-Guided Dialogue dataset (Rastogi et al., 2020) – showing a genuine move beyond English. Translating existing corpora offers a pragmatic way to extend coverage to low-resource languages but comes with caveats. Conversation reflects culture: idioms, politeness norms, turn-taking cues, and relevance judgments are culturally embedded and may not transfer cleanly through translation (Zhang and Toral, 2019; Majewska et al., 2023). Automatic translation can aid scalability but should always involve native speakers – not merely as reviewers, but as active collaborators in the process.

Most created datasets (∼61%) involve pairs of participants in short- or long-term conversations. Some consist of adjacent utterance pairs extracted

[3]The last ranking was made in 2023 and LREC has presumably risen since then.

[4]Findings are not indexed in the ICORE Conference Portal but are considered less prominent than main conferences.

from longer dialogues with potentially more speakers. Three datasets feature imbalanced roles, with one main speaker and several sub-speakers or listeners, including one study where a single person enacts multiple speakers in virtual reality (Lai et al., 2024). About 13% of the datasets include more than two speakers: in some cases, the number is explicitly defined and the study is designed around multilogues, while in others, multi-party dynamics emerge incidentally.

Written data remains the dominant modality (∼57%), but multimodal datasets also represent a substantial share (∼28%). These include collections combining written conversations with related video or speech (e.g. YouTube comments), as well as transcribed dialogues from videos or audio recordings, sometimes enriched with additional data such as eye-tracking or physiological signals.

Nearly half of the created datasets contain human-produced data with varying levels of interaction and freedom. This includes 24% of human–human dialogues and 10% of online conversations, such as social media exchanges or YouTube comments, representing the most interactive and least constrained forms. Another 7% consist of dialogues from TV shows, which approximate spontaneous conversation to different degrees – scripts being generally less naturalistic than actor transcripts. Additionally, 8% rely on hand-crafted dialogues, written by humans following specific instructions, similar to screenwriters. However, real conversations include backchannels, disfluencies, overlaps, and other interactional phenomena often absent from fictional dialogues (Chepinchikj and Thompson, 2016). Such dynamics are crucial for mutual understanding but inconvenient for readers, so naturalistic transcripts are rare, complicating the creation of human-like scripted dialogues (Pilan et al., 2024). While this is acceptable when conversational dynamics are not under study, ensuring that degraded dialogue quality does not bias other analyses requires careful evaluation.

Similarly, about one third of the created datasets were generated by LLMs. These are much easier to produce than datasets involving real human interactions, which are costly and time-consuming, especially given current data demands. However, LLM-generated conversations should not be considered equivalent to human–human ones. Their creation settings impose structural constraints: LLMs typically produce clean, alternating turns, unlike the irregular patterns of real dialogue. Moreover, they are mostly written, despite aiming to mimic spontaneous conversation. Crucially, LLMs "learn" and generate language in ways fundamentally different from humans (Bender and Koller, 2020); lacking grounding, interaction, and shared attention, their outputs cannot reliably represent human communicative behaviour, and thus corpora generated in this way cannot serve as reliable evidence for studying how people communicate. Nonetheless, such resources can support analytical work comparing human and machine dialogue. Three created datasets included both types to study their differences (Li et al., 2024b; Occhipinti et al., 2024) or improve LLMs (Sato et al., 2024). Human–machine corpora also account for about 9%, consistent with the prevalence of DSs. However, the strong research focus on these technologies warrants reflection. Access to DSs is not universal, and they should not be seen as substitutes for human contact, particularly for isolated populations such as the elderly. Assessing actual user needs remains essential when developing such tools (Kann et al., 2022).

The rest of the created datasets contained no conversation except for one which consists of non-player characters dialogues from a video game (Weir et al., 2024).

7% of the surveyed papers extended existing datasets with human-made and/or synthetic annotations. A few papers (about 4%) also present generation methods for producing synthetic data.

Five papers present benchmarks for evaluating LLMs, each targeting different aspects – from specific elements like social norm incorporation to broader traits such as human-likeness or multi-turn ability. They cover various languages: two address English and Chinese (Duan et al., 2024; Ou et al., 2024), one focuses solely on Chinese (Zhan et al., 2024), one on Korean (Jang et al., 2024), and one omits language specification (Li et al., 2024a) but is based on the English Wizards of Wikipedia (Dinan et al., 2019). While evaluating LLMs is essential given their growing prominence, releasing fixed benchmarks risks rapid obsolescence: they suit pre-existing models but may later enter training data, compromising validity. Similarly, benchmarks derived from existing corpora can introduce bias. Nonetheless, producing evaluation datasets in languages other than English may yield a positive side effect – enriching resources for under-represented languages and promoting greater linguistic diversity in future LLMs.

5.2. Used Datasets

Regarding usage, the most frequently employed corpora in the reviewed papers were variants of MultiWOZ (Budzianowski et al., 2018; Eric et al., 2020; Zang et al., 2020), unsurprising given the high proportion of work on TOD systems. These corpora contain task-oriented human–human chat-based dialogues collected in an adaptation of the Wizard-of-Oz (WOZ) setting, where one participant acts as a customer performing a task (e.g. booking a restaurant), and the other participant has access

to a database to respond, which mimics what the system would have access to. In the original WOZ setting, the first participant was told that they were talking to a machine and the second was expected to play the system, which is not the case in MultiWOZ.

Originally in English, MultiWOZ has since been translated into several languages. While reusing such established corpora is methodologically convenient, recent LLM-based approaches make this practice increasingly problematic. Because these datasets are used for both training and evaluation (see section 6), building a TOD system often reduces to adapting an LM to reproduce expected WOZ outputs rather than generalising beyond them. Moreover, these corpora were likely included in LLM training data, reinforcing data circularity and undermining the reliability of ground-truth–based evaluations.

A broad range of other corpora are also employed across different studies. Among the most frequently cited are DailyDialog (Li et al., 2017) and SGD (Rastogi et al., 2020; Lee et al., 2022) – another TOD corpus. Zahiri and Choi (2018) and Poria et al. (2019), based on the TV-show *Friends* and IEMOCAP (Busso et al., 2008) – an acted multimodal multi-speaker corpus – are widely used for conversation emotion recognition. Various datasets are used to enhance different aspects of DSs: PersonaChat (Zhang et al., 2018) – persona-based conversations in pairs – for more personalised chit-chat systems, ESConv – emotional support conversations (Liu et al., 2021) between two crowdworkers – for building emotional support DSs, ReDial (Li et al., 2018) – movie recommendation dialogues – for recommender systems, QReCC (Anantha et al., 2021) – hand-crafted dialogues with QA pairs grounded in web pages – for QA, and DSTC (Anantha et al., 2021) – phone calls between bus passengers and DSs – for DST. All of these most used datasets are in English.

Both in terms of used and created resources, a recurring issue is the lack of information w.r.t. the datasets used. Many papers omit crucial details such as language, dialogue type (handcrafted, scripted, or natural), or participant configuration (human–human vs. human–machine). This lack of transparency affects both reproducibility and model validity. For example, although DailyDialog is often described as everyday conversation, its dialogues were handcrafted, not collected from real interactions. While such data can be high-quality, it reflects our *representation* of conversation rather than authentic exchanges, often underestimating features like pauses, repairs, and disfluencies – core markers of spoken dialogue. This highlights a broader need to clarify which conversational features are desirable in models, as these depend on task objectives. For instance, hesitations may enhance human-likeness but their necessity in TOD systems can be questioned, unless they significantly improve information processing. Balancing such trade-offs calls for coordinated investigation within the community.

6. Evaluation

Our findings confirm that conversational AI remains the central focus of current research. Consequently, we concentrate on the categories Generative Systems and LM Evaluation defined in section 4 to assess prevailing evaluation practices within the community.

Two aspects are relevant when it comes to conversational generative systems: their ability to tackle the specific task they were designed for, which lies outside the scope of this paper, and their conversational competency.

Regarding conversational evaluation, nearly 45% of the studies rely on reference-based automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). These metrics compare system outputs to gold-standard utterances based on lexical overlap. Although convenient and widely used, they fail to capture the interactive and context-sensitive nature of dialogue, often penalising diverse yet valid responses – a defining feature of natural conversation. Notably, about one third of studies use no additional evaluation criteria, which is problematic given that conversational quality is multi-faceted and cannot be reduced to semantic similarity with a ground truth. Such limited assessments hinder our ability to determine whether a DS is genuinely useful or superior to simpler baselines.

The increasing complexity of conversational tasks has encouraged the use of LLM-based evaluators, capable of flexibly scoring multiple dialogue dimensions. Among DS papers, about 12% rely on such evaluations[5]. However, over one fourth do not validate LLM judgments with human input, even though LLMs may have divergent "expectations" of conversational quality in relation to humans. Furthermore, many corpora used for training and evaluation likely appear in the LLMs' training data, biasing the results. Likewise, models are often assessed on benchmarks already seen during training, raising concerns about data leakage and evaluation validity.

In human evaluations, the most assessed dimensions are coherence (18% of DS papers with human evaluation) and fluency (16%). Other aspects such as relevance, consistency, and naturalness

[5]We exclude BERTScore (Zhang et al., 2019), focusing instead on prompt-based approaches using models like Llama, Mistral, or ChatGPT.

are also examined (each 8–9%), highlighting the need to decompose evaluation into more objective features assessable by external raters. Overall conversational quality is also considered (9%), though its subjective nature makes it difficult to measure consistently. Notably, very few systems are tested by end users in interaction ($\leq$ 6%), even though extrinsic evaluations may fail to reflect real-world experience.

Taken together, these observations suggest that without careful reconsideration of evaluation practices, the field risks measuring progress in ways that may be convenient but ultimately misleading, reinforcing the need for more rigorous and context-aware assessment methods.

7. Discussion

Data at the Core of Dialogue Research The current trajectory of dialogue research reveals a tendency to prioritise convenience over suitability, particularly in the selection of data and evaluation methods. Data should be at the heart of conversational research – and NLP/CL research in general, yet it is too often treated as a tool to train models rather than an object of careful study in its own right. While few corpora capture spontaneous, face-to-face interaction due to practical constraints, not all proxies are equivalent. Hand-crafted dialogues, movie scripts, human–machine exchanges, and LLM-generated conversations each reflect different interactional dynamics and limitations. Recognising and explicitly defining these limits is crucial. A precise understanding of a corpus's scope should precede model development or evaluation, ensuring alignment between data and research goals.

Evaluation Beyond Convenience Evaluation practices present another significant challenge. Systems without rigorous evaluation are meaningless, yet the field often relies on metrics chosen for convenience or historical precedent, despite their known limitations. Automatic evaluation cannot replace careful, task-specific assessment, and the cost of human evaluation should not justify inadequate metrics. Instead, we must clearly define system expectations and decompose them into measurable aspects. Similarly, using evaluation proxies for humans must be done cautiously and cannot be generalised across tasks; validation in one context does not ensure validity in another. Without human validation, reliance on LLM-based evaluation risks creating a self-reinforcing evaluation loop, particularly when the evaluation corpora themselves may have been included in the models' training data. Such circularity can overestimate system performance and obscure meaningful limitations.

Limitations of Generative Dialogue Generative systems, despite their fluency, have been shown to diverge from human interaction (Ivey et al., 2024; Zhang and Yu, 2025). Their conversations are rigidly turn-based, constrained by designer-defined intents, and optimised for engagement or utility rather than authenticity. Unlike humans, they lack interactional dynamics, and even if they could perfectly mimic surface-level conversational features, they would still fail to capture the grounded, embodied, and socially situated nature of human dialogue (Bender et al., 2021).

Rethinking Human-Likeness The question of human-likeness in dialogue further illustrates these limitations. LLMs can sometimes pass restricted forms of the Turing test (Jones and Bergen, 2025), yet evidence shows that evaluators may prefer machine-generated outputs over human dialogue in controlled settings. This does not imply that models are indistinguishable from humans. Rather, it highlights the unreliability of human judgments of conversational human-likeness, particularly under novel or brief evaluation conditions. The emergence of LLMs has also shifted public perception: whereas face-to-face interactions rarely prompt doubt about human interlocutors, non-face-to-face interactions with unknown partners are now more often approached with scepticism. Real human-human conversations are noisy, filled with overlaps, hesitations, repairs, and subtle cues that often go unnoticed by the participants. We are indeed not frequently exposed to conversation from an outside perspective. While repairs and backchannels are crucial for efficient interaction, they do not need to be consciously remembered in order to follow the conversation. Thus, showing a transcript *a posteriori* to people and asking them to rate it for human-likeness will probably not reflect how they would have rated it *during* the conversation. Even placing someone in the role of a judge while they are having a conversation could change the way they interact and make them notice behaviours they would not have consciously analysed normally, making them doubt the quality of the conversation. Paradoxically, interactional patterns may possibly penalise authentic interactions when they are evaluated externally, while LLMs would produce a smoother dialogue.

Rethinking Goals for Dialogue Systems These reflections call for a reconsideration of what DSs are meant to achieve. The goal should not necessarily be human imitation, which carries risks of manipulation, misinformation, and emotional over-attachment. Instead, systems should be evaluated on their ability to perform specific tasks effectively. But identifying which conversational features support function requires a strong theoretical

foundation. A formal characterisation of dialogue would support the design of dimension-specific metrics, clarify the distinction between human-human and human-machine interaction, and better align datasets, models, and evaluations with explicit communicative objectives.

8. Conclusion

The rise of LLMs developed by major technology companies in opaque ways raises fundamental questions about how we develop and evaluate conversational systems. Since reproducing LLMs at this scale is not feasible, we must seek alternative ways to regain control over our work. The solution, however, cannot be to restrict access to data or reverse the progress made toward open science. Instead, the community must focus on research methodology: clearly defining tasks, thoroughly characterising datasets, and designing meaningful, task-specific evaluation metrics.

A key priority is to make data creation more efficient in terms of time and resources. Developing and improving tools that support human language collection, and, for the specific focus of this paper, human *dialogical* interactions, should become a central concern of our community if we hope to overcome the evaluation challenges posed by LLMs. This endeavour must foreground ethical considerations regarding acceptable data collection methods and purposes. LLM-generated dialogues should *not* become substitutes for human-produced resources. Instead, ongoing efforts to build multilingual, multimodal, and low-resource language corpora should be further encouraged.

To ensure transparency and comparability in dialogue research, datasets should be systematically described along a few core dimensions:

(1) Speakers and Language Who are the participants, and how many? What languages, dialects, or sociolinguistic varieties are represented? Are the speakers native or L2 users, or is the material a translation? What social dynamics shape interactions, and, are these preserved or distorted through translation or simulation?

(2) Interaction and Medium Through what medium did the interaction occur, and how is it characterised in terms of modalities and synchrony? How are interactional mechanisms affected by the medium? What is captured or lost between the conversation and the medium of study?

(3) Task and Evaluation Context What communicative goal or task underlies the conversation, and how well does it align with the study objectives? Which specific conversational dimensions are targeted by the evaluation metrics, and which are irrelevant to the task? Has the evaluation corpus been confirmed as unseen by the model?

This list of dimensions is not exhaustive and should be viewed as a starting point towards a systematic framework for describing dialogue data. It must be adapted to the nature of the research – whether the aim is to build practical systems, analyse human interaction, or develop theoretical models. A formal typology, capable of capturing variation across modalities, interaction types, and communicative goals, would help us ensure that evaluation data are representative and comparable across studies, keeping dialogue research interpretable, reproducible, and meaningful in the era of LLMs.

9. Ethics Statement / Broader Impact

Natural Language Processing is by Nature multidisciplinary and keeping track of the numerous theories in each field is a colossal task. Moreover, the pressure to publish is intense and even more so for younger researchers without the safety of a permanent position. Thus the purpose of this study is not to highlight specific papers where questionable research practices have been observed as all of us could be guilty of such lapses at some point in our careers. On the contrary, our aim is to raise awareness on widespread but often overlooked practices in the field and to encourage the conversation between different sub-communities of NLP/CL. For this reason, we did not point to specific references when discussing bad practices we identified in the ACL Anthology 2024.

10. Limitations

Since this study was based solely on papers extracted from the ACL Anthology 2024, we did not include many NLP/CL venues, particularly journals, which are often a more suitable venue for longer-form research such as detailed analysis and modelling. However we believe that these types of studies also have a place in conferences, as they should form the conceptual backbone of the models presented there.

11. Acknowledgments

We thank the anonymous reviewers for their valuable feedback.

12. Bibliographical References

Yann Algan, Sergei Guriev, Elias Papaioannou, and Evgenia Passari. 2017. The european trust crisis and the rise of populism. Brookings papers on economic activity, 2017(2):309–400.

Maxime Amblard, Karën Fort, Michel Musiol, and Manuel Rebuschi. 2014. L'impossibilité de l'anonymat dans le cadre de l'analyse du discours. In Journée ATALA éthique et TAL, Paris, France.

Satanjeev Banerjee 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, Ann Arbor, Michigan. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. Linguistic Issues in Language Technology, 6.

Emily M. Bender, 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.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198, Online. Association for Computational Linguistics.

Giovanni Maria Biancofiore, Yashar Deldjoo, Tommaso Di Noia, Eugenio Di Sciascio, and Fedelucio Narducci. 2024. Interactive question answering systems: Literature review. ACM Comput. Surv., 56(9).

Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. 2025. A survey of recent advances on turn-taking modeling in spoken dialogue systems. In Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 254–271, Bilbao, Spain. Association for Computational Linguistics.

Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. 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.

Neda Chepinchikj and Celia Thompson. 2016. Analysing cinematic discourse using conversation analysis. Discourse, Context & Media, 14:40–53.

Herbert H. Clark. 1996. Using language. Cambridge university press.

Herbert H. Clark and Edward F. Schaefer. 1992. Dealing with overhearers. Arenas of language use, pages 248–274.

Guillaume Dezecache and Robin Dunbar. 2012. Sharing the joke: The size of natural laughter groups. Evolution and Human Behavior, 33:775–779.

Mark Dingemanse, Seán G. Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S. Gisladottir, Kobin H. Kendrick, Stephen C. Levinson, Elizabeth Manrique, Giovanni Rossi, and N. J. Enfield. 2015. Universal principles in the repair of communication problems. PLOS ONE, 10(9):1–15.

Fanny Ducel, Karën Fort, Gaël Lejeune, and Yves Lepage. 2022. Langues par défaut ? analyse contrastive et diachronique des langues non citées dans les articles de TALN et d'ACL (contrastive and diachronic study of unmentioned (by default ?) languages in TALN and ACL we study the application of the #BenderRule in natural language processing articles, taking into account a contrastive and a diachronic dimensions, by examining the proceedings of two NLP conferences, TALN and ACL, over time). In Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, pages 144–153, Avignon, France. ATALA.

R. I. M. Dunbar, N. D. C. Duncan, and D. Nettle. 1995. Size and structure of freely forming conversational groups. Human Nature, 6(1):67–78.

Xiachong Feng, 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.

Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. 2016. Semantic annotation of the ACL Anthology corpus for the automatic analysis of scientific literature. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3694–3701, Portorož, Slovenia. European Language Resources Association (ELRA).

Yi-Li Hsu, Jui-Ning Chen, Yang Fan Chiang, Shang-Chien Liu, Aiping Xiong, and Lun-Wei Ku. 2024. Enhancing perception: Refining explanations of news claims with LLM conversations. In Findings of the Association for Computational Linguistics: NAACL 2024, pages 2129–2147, Mexico City, Mexico. Association for Computational Linguistics.

Takeru Isaka, Atsushi Otsuka, and Iwaki Toshima. 2024. Analysis of sensation-transfer dialogues in motorsports. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 876–886, Torino, Italia. ELRA and ICCL.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, Dustin Wright, Abraham Israeli, Anders Giovanni Møller, Lechen Zhang, and David Jurgens. 2024. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue.

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. 2022. "do you follow me?": A survey of recent approaches in dialogue state tracking. In Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 336–350, Edinburgh, UK. Association for Computational Linguistics.

Gail Jefferson. 1972. Side sequences. In David Sudnow, editor, Studies in Social Interaction, chapter 9, page 294–338. Free Press, New York.

Cameron R. Jones and Benjamin K. Bergen. 2025. Large language models pass the turing test.

Jose Joskowicz. 2023. Video conferencing technologies: Past, present and future.

Cristóbal Rovira Kaltwasser and Paul Taggart. 2025. Populism and democracy: The road ahead. PS: Political Science & Politics, 58(1):96–100.

Katharina Kann, Shiran Dudy, and Arya D. McCarthy. 2022. A major obstacle for NLP research: Let's talk about time allocation! In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8959–8969, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kobin H. Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. Philosophical transactions of the royal society B, 378(1875):20210473.

Jaimie Arona Krems, Robin I. M. Dunbar, and Steven Neuberg. 2016. Something to talk about: are conversation sizes constrained by mental modeling abilities? Evolution and Human Behavior, 37(6):423–428. Funding Information: RD's research is funded by a European Research Council Advanced grant (295663). We thank Alexander Danvers and Sean C. Murphy for their helpful comments on earlier versions of this work. We thank Jackie Swift-Honer for her help formatting this article for submission. Publisher Copyright: © 2016 Elsevier Inc.

Jaimie Arona Krems and Jason Wilkes. 2019. Why are conversations limited to about four people? a theoretical exploration of the conversation size constraint. Evolution and Human Behavior, 40(2):140–147.

Sheetal Kusal, 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.

Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. 2023. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. Machine Intelligence Research, 20(3):318–334.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. Science, 359(6380):pp. 1094–1096.

Wonjun Lee, San Kim, and Gary Geunbae Lee. 2024. Enhancing dialogue speech recognition with robust contextual awareness via noise representation learning. In Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 333–343, Kyoto,

Japan. Association for Computational Linguistics.

Andrew Li, Zhenduo Wang, Ethan Mendes, Duong Minh Le, Wei Xu, and Alan Ritter. 2024. ChatHF: Collecting rich human feedback from real-time conversations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 270–279, Miami, Florida, USA. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022. A survey of discourse parsing. Frontiers of Computer Science, 16(5):165329.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

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

Andy Lücking, Felix Voll, Daniel Rott, Alexander Henlein, and Alexander Mehler. 2025. Head and hand movements during turn transitions: Data-based multimodal analysis using the Frankfurt VR Gesture–Speech Alignment Corpus (FraGA). In Proceedings of the 29th Workshop on The Semantics and Pragmatics of Dialogue, SemDial'25 – Biaogue. Forthcoming.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M. Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. Transactions of the Association for Computational Linguistics, 11:139–156.

Lorenza Mondada. 2019. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. Journal of Pragmatics, 145:47–62. Quo Vadis, Pragmatics?

Pippa Norris and Ronald Inglehart. 2019. Cultural Backlash: Trump, Brexit, and Authoritarian Populism. Cambridge University Press.

Sanusi Bernice Oluwalanumi Omipidan, Ismail Adewale. 2024. Rise of social media in the digital age: Whatsapp a threat to effective communication. IMSU Journal of Communication Studies, 8(1).

Elisa Omodei, Yufan Guo, Jean-Philippe Cointet, and Thierry Poibeau. 2014. Argumentative analysis of the ACL Anthology (analyse argumentative du corpus de l'ACL (ACL Anthology)) [in French]. In Proceedings of TALN 2014 (Volume 2: Short Papers), pages 580–585, Marseille, France. Association pour le Traitement Automatique des Langues.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ildiko Pilan, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2024. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. In Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 440–457, Kyoto, Japan. Association for Computational Linguistics.

Aniket Pramanick, Yufang Hou, Saif Mohammad, and Iryna Gurevych. 2023. A diachronic analysis of paradigm shifts in NLP research: When, how, and why? In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2312–2326, Singapore. Association for Computational Linguistics.

Sanjana Ramprasad, Elisa Ferracane, and Zachary Lipton. 2024. Analyzing LLM behavior in dialogue summarization: Unveiling circumstantial hallucination trends. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12549–12561, Bangkok, Thailand. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In Jim Schenkein, editor, Studies in the Organization of Conversational Interaction, pages 7–55. Academic Press.

Emanuel A. Schegloff. 1979. The Relevance of Repair to Syntax-for-Conversation, pages 261 – 286. Brill, Leiden, The Netherlands.

Emanuel A. Schegloff. 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In Deborah Tannen, editor, Analyzing Discourse: Text and Talk, page 71–93. Georgetown University Press, Washington, D.C.

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

David Traum. 2003. Issues in multiparty dialogues. In Workshop on Agent Communication Languages, pages 201–211. Springer.

Jacqueline van Arkel, Marieke Woensdregt, Mark Dingemanse, and Mark Blokpoel. 2020. A simple repair mechanism can alleviate computational demands of pragmatic reasoning: simulations and complexity analysis. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 177–194, Online. Association for Computational Linguistics.

Vijesh Vijayan, Temin Thomas, and Princy D Nellanat. 2025. Mapping fake news and misinformation in media: A two-decade bibliometric review of emerging trends. Insight-News Media, 8(1):734–734.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL Anthology. In Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. 2024. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15880–15893, Bangkok, Thailand. Association for Computational Linguistics.

Jing Yang, Nils Feldhus, Salar Mohtaj, Leonhard Hennig, Qianli Wang, Eleni Metheniti, Sherzod Hakimov, Charlott Jakob, Veronika Solopova, Konrad Rieck, David Schlangen, Sebastian Möller, and Vera Schmitt. 2026. Order in the evaluation court: A critical analysis of nlg evaluation trends.

Fulei Zhang and Zhou Yu. 2025. Mind the gap: Linguistic divergence and adaptation strategies in human-llm assistant vs. human-human interactions. Presented at the Workshop on Generative AI for E-Commerce, at RecSys 2025.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pages 73–81, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

13. Language Resource References

Anantha, Raviteja and Vakulenko, Svitlana and Tu, Zhucheng and Longpre, Shayne and Pulman, Stephen and Chappidi, Srinivas. 2021. Open-Domain Question Answering Goes Conversational via Question Rewriting. Association for Computational Linguistics. PID https://github.com/apple/ml-qrecc.

Arana, Janire and Idoyaga, Mikel and Urruela, Maitane and Espina, Elisa and Atutxa Salazar, Aitziber and Gojenola, Koldo. 2024. A Virtual Patient Dialogue System Based on Question-Answering on Clinical Records. ELRA and ICCL. PID https://github.com/Midoiaga/VirPat-2024.

Budzianowski, Paweł and Wen, Tsung-Hsien and Tseng, Bo-Hsiang and Casanueva, Iñigo and Ultes, Stefan and Ramadan, Osman and Gašić, Milica. 2018. MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. Association for Computational Linguistics. PID https://github.com/budzianowski/multiwoz/blob/master/data/MultiWOZ_1.0.zip.

Busso, Carlos and Bulut, Murtaza and Lee, Chi-Chun and Kazemzadeh, Abe and Mower, Emily and Kim, Samuel and Chang, Jeannette N. and Lee, Sungbok and Narayanan, Shrikanth S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Springer. PID https://sail.usc.edu/iemocap/.

Bychkova, Polina and Yaskevich, Alyaxey and Gyulasaryan, Serafima and Rakhilina, Ekaterina. 2024. Building a Database of Conversational Routines. ELRA and ICCL.

Emily Dinan and Stephen Roller and Kurt Shuster and Angela Fan and Michael Auli and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. ICLR. PID https://parl.ai/projects/wizard_of_wikipedia/.

Duan, Haodong and Wei, Jueqi and Wang, Chonghua and Liu, Hongwei and Fang, Yixiao and Zhang, Songyang and Lin, Dahua and Chen, Kai. 2024. BotChat: Evaluating LLMs'

Capabilities of Having Multi-Turn Dialogues. Association for Computational Linguistics. PID https://github.com/open-compass/BotChat.

Eric, Mihail and Goel, Rahul and Paul, Shachi and Sethi, Abhishek and Agarwal, Sanchit and Gao, Shuyang and Kumar, Adarsh and Goyal, Anuj and Ku, Peter and Hakkani-Tur, Dilek. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. European Language Resources Association. PID https://github.com/budzianowski/multiwoz/blob/master/data/MultiWOZ_2.1.zip.

Htun, Hay Man and Kyaw Thu, Ye and Chanlekha, Hutchatai and Funakoshi, Kotaro and Supnithi, Thepchai. 2024. myMediCon: End-to-End Burmese Automatic Speech Recognition for Medical Conversations. ELRA and ICCL. PID https://github.com/ye-kyaw-thu/myMediCon.

Jang, Seongbo and Lee, Seonghyeon and Yu, Hwanjo. 2024. KoDialogBench: Evaluating Conversational Understanding of Language Models with Korean Dialogue Benchmark. ELRA and ICCL. PID https://github.com/sb-jang/kodialogbench.

Lai, Viet Dac and Pham, Duy Ngoc and Steinberg, Jonathan and Mikeska, Jamie and Nguyen, Thien Huu. 2024. CAMAL: A Novel Dataset for Multi-label Conversational Argument Move Analysis. ELRA and ICCL.

Lee, Harrison and Gupta, Raghav and Rastogi, Abhinav and Cao, Yuan and Zhang, Bin and Wu, Yonghui. 2022. SGD-X: A Benchmark for Robust Generalization in Schema-Guided Dialogue Systems. AAAI. PID https://github.com/google-research-datasets/dstc8-schema-guided-dialogue.

Li, Raymond and Kahou, Samira and Schulz, Hannes and Michalski, Vincent and Charlin, Laurent and Pal, Chris. 2018. Towards deep conversational recommendations. Curran Associates Inc., NIPS'18. PID https://redialdata.github.io/website/.

Li, Xiangci and Song, Linfeng and Jin, Lifeng and Mi, Haitao and Ouyang, Jessica and Yu, Dong. 2024a. A Knowledge Plug-and-Play Test Bed for Open-domain Dialogue Generation. ELRA and ICCL. PID https://github.com/jacklxc/Ms.WoW.

Li, Yanran and Su, Hui and Shen, Xiaoyu and Li, Wenjie and Cao, Ziqiang and Niu, Shuzi. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. Asian Federation of Natural Language Processing. PID http://yanran.li/dailydialog.html.

Li, Yu and Hazarika, Devamanyu and Jin, Di and Hirschberg, Julia and Liu, Yang. 2024b. From Pixels to Personas: Investigating and Modeling Self-Anthropomorphism in Human-Robot Dialogues. Association for Computational Linguistics. PID https://github.com/yooli23/Pix2Persona.

Liu, Siyang and Zheng, Chujie and Demasi, Orianna and Sabour, Sahand and Li, Yu and Yu, Zhou and Jiang, Yong and Huang, Minlie. 2021. Towards Emotional Support Dialog Systems. Association for Computational Linguistics. PID https://github.com/thu-coai/Emotional-Support-Conversation.

Occhipinti, Daniela and Tekiroğlu, Serra Sinem and Guerini, Marco. 2024. PRODIGy: a PROfile-based DIalogue Generation dataset. Association for Computational Linguistics. PID https://github.com/LanD-FBK/prodigy-dataset.

Ou, Jiao and Lu, Junda and Liu, Che and Tang, Yihong and Zhang, Fuzheng and Zhang, Di and Gai, Kun. 2024. DialogBench: Evaluating LLMs as Human-like Dialogue Systems. Association for Computational Linguistics. PID https://github.com/kwai/DialogBench.

Poria, Soujanya and Hazarika, Devamanyu and Majumder, Navonil and Naik, Gautam and Cambria, Erik and Mihalcea, Rada. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. Association for Computational Linguistics, ISLRN 190-235-172-195-7.

Rastogi, Abhinav and Zang, Xiaoxue and Sunkara, Srinivas and Gupta, Raghav and Khaitan, Pranav. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. AAAI. PID https://github.com/google-research-datasets/dstc8-schema-guided-dialogue.

Sato, Shiki and Akama, Reina and Suzuki, Jun and Inui, Kentaro. 2024. A Large Collection of Model-generated Contradictory Responses for Consistency-aware Dialogue Systems. Association for Computational Linguistics. PID https://github.com/shiki-sato/rgm-contradiction.

Tu, Lifu and Qu, Jin and Yavuz, Semih and Joty, Shafiq and Liu, Wenhao and Xiong, Caiming and Zhou, Yingbo. 2024. Efficiently Aligned Cross-Lingual Transfer Learning for Conversational Tasks using Prompt-Tuning. Association for Computational Linguistics. PID https://console.cloud.google.com/storage/browser/multilingual-sgd-data-research.

Wang, Junda and Yao, Zonghai and Yang, Zhichao and Zhou, Huixue and Li, Rumeng and Wang, Xun and Xu, Yucheng and Yu, Hong. 2024. NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes. Association for Computational Linguistics. PID https://github.com/believewhat/Dr.NoteAid.

Weir, Nathaniel and Thomas, Ryan and d'Amore, Randolph and Hill, Kellie and Van Durme, Benjamin and Jhamtani, Harsh. 2024. Ontologically Faithful Generation of Non-Player Character Dialogues. Association for Computational Linguistics. PID https://github.com/nweir127/KNUDGE.

Zahiri, Sayyed M. and Choi, Jinho D. 2018. Emotion Detection on TV Show Transcripts with Sequence-Based Convolutional Neural Networks. AAAI Workshops.

Zang, Xiaoxue and Rastogi, Abhinav and Sunkara, Srinivas and Gupta, Raghav and Zhang, Jianguo and Chen, Jindong. 2020. MultiWOZ 2.2 : A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. Association for Computational Linguistics. PID https://github.com/budzianowski/multiwoz/blob/master/data/MultiWOZ_2.2.

Zhan, Haolan and Li, Zhuang and Kang, Xiaoxi and Feng, Tao and Hua, Yuncheng and Qu, Lizhen and Ying, Yi and Chandra, Mei Rianto and Rosalin, Kelly and Jureynolds, Jureynolds and Sharma, Suraj and Qu, Shilin and Luo, Linhao and Zukerman, Ingrid and Soon, Lay-Ki and Semnani Azad, Zhaleh and Haf, Reza. 2024. RENOVI: A Benchmark Towards Remediating Norm Violations in Socio-Cultural Conversations. Association for Computational Linguistics. PID https://github.com/zhanhl316/ReNoVi.

Zhang, Saizheng and Dinan, Emily and Urbanek, Jack and Szlam, Arthur and Kiela, Douwe and Weston, Jason. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? Association for Computational Linguistics. PID https://www.kaggle.com/datasets/atharvjairath/personachat.