Conversational Grounding in Large Language Models: Evaluation Methods, Challenges and Future Directions

Michelle Elizabeth[1, 2], **Gwénolé Lecorvé**[1], **Lina Rojas-Barahona**[1], **Magalie Ochs**[2]
[1]Orange Research, [2]LIS, Aix-Marseille Université
Correspondence: michelle.elizabeth@orange.com

Abstract

Conversational grounding is the collaborative process through which speakers establish and maintain mutual understanding. It is essential for the success of a dialogue. While it is inherent in human conversations, it remains a challenge for instruction-following Large Language Models (LLM). This paper surveys how conversational grounding is evaluated in task-oriented dialogue in the current era of LLMs. First, we focus on how conversational grounding is modelled explicitly—using dialogue acts and by modelling the participant mental state. Then, we review collaborative tasks that enable the evaluation of conversational grounding implicitly at the global level based on outcomes. Finally, we highlight the limitations of evaluation, notably the current methodology and metrics used, and outline research directions in conversational grounding and its evaluation.

1 Introduction

In human dialogue, the participants provide explicit cues of their degree of understanding via acknowledgments, back-channels, paraphrasing, clarifications as well as repair and recover from errors. This process of collaboratively building a mutual understanding or common ground is known as *conversational grounding* (Clark and Schaefer, 1989). Unlike human-human dialogue, grounding is not inherent in human-machine dialogue. Despite their fluent responses, Large Language Models (LLMs) make very little effort in clarifying ambiguities or asking for more relevant information to provide informed responses (Shaikh et al., 2024, 2025; Wang et al., 2025a). In Task-Oriented Dialogue (TOD), failure to ground can lead agents to provide confident responses based on incomplete information, which impacts the ability to achieve the user goal (Elizabeth et al., 2025). In critical domains like healthcare, grounding is essential as even the tiniest misunderstanding of the user request can have serious consequences (Wang et al., 2025b; Foo et al., 2025). Hence, it is necessary to develop reliable evaluation methods for conversational grounding alongside grounding strategies.

Traditional dialogue pipelines were designed to handle uncertainty in the input and to ask for user feedback when required (Su et al., 2016). In contrast, LLMs are tuned to directly follow instructions, which impacts grounding behaviours like clarifications and repairs (Shaikh et al., 2025). Grounding evaluates beyond surface-level response quality by analyzing semantic and pragmatic levels through dialogue phenomena like coreference, anaphora, etc. (Mohapatra et al., 2024b). To ensure that dialogue systems truly understand user goals, it is essential to evaluate *how and to what extent these systems are able to ground.* Hence, in this paper, we survey the evaluation of conversational grounding for TOD in key papers from the recent LLM-based approaches in the literature.

The remainder is organized as follows. Section 2 first positions our work by distinguishing between conversational grounding and another type of grounding called *symbol grounding*, as well as differentiating our survey from other previous comparable studies. Then, Section 3 lays the theoretical background about grounding as studied in cognitive science, and Section 4 organises the recent LLM-based TOD literature focusing on conversational grounding, i.e., how grounding is modelled and which evaluation strategies are adopted. Finally, Section 5 presents the research directions that we believe emerge from this literature review.

2 Positioning of our Work

This section clarifies the scope of our survey by contrasting conversational grounding with the complementary concept of symbol grounding, and highlights the novelty and contributions of our work relative to existing surveys.

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 151–163
August 2–5, 2026.

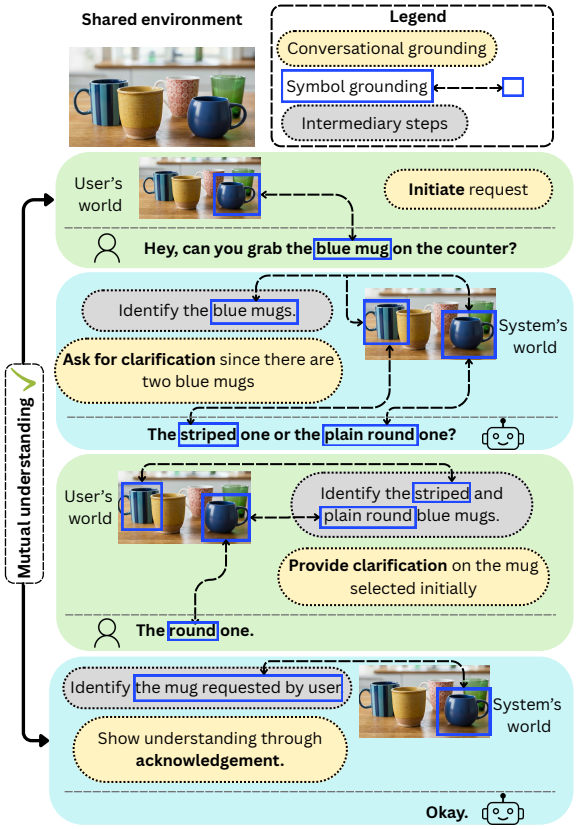


Figure 1: An example showing the presence of conversational grounding and symbol grounding in dialogue.

2.1 Symbol versus Conversational Grounding

Grounding is widely used in dialogue literature. Hence, it is important to distinguish between the two types of grounding. The first one is *symbol grounding*, which concerns the link between language and entities present in the participants' *world*. These entities can be objects in the physical scene or in an image, facts or concepts from the participant's knowledge, etc. (Larsson, 2018; Benotti and Blackburn, 2021). Specialised areas of research under the umbrella of symbol grounding thus include visual grounding (Xiao et al., 2025) and knowledge grounding (Schneider et al., 2024). Symbol grounding is also known as static grounding (Chandu et al., 2021) and perceptual grounding (Schneider et al., 2024). It can be evaluated by measuring whether the external entity is correctly used in the dialogue.

The second type of grounding—and the focus of the paper—is *conversational grounding*. It studies the interactive process for establishing mutual understanding in dialogue (Schneider et al., 2024; Mohapatra et al., 2024a). It is also known as communicative grounding (Larsson, 2018) and collaborative grounding (Benotti and Blackburn, 2021). The evaluation is driven by the presence of evidence of grounding behaviour (e.g. acknowledgements, clarification, correction).

Both symbol and conversational grounding are essential to form a complete understanding of the conversational context, as shown in Figure 1. Symbol grounding is essential within conversational grounding in order to make sense of words and their real-world counterparts. This matters in evaluation as well, since systems can try to establish mutual understanding without understanding the physical entity and vice versa. In this paper, we focus specifically on evaluating *conversational grounding*.

2.2 Related Surveys

A majority of the studies on grounding deal with symbol grounding. Chandu et al. (2021) found that *grounding* was used in the symbolic sense in the pre-ChatGPT (Ouyang et al., 2022) phase when LLMs were not mainstream. They emphasise the need for more focus on the process of conversational grounding. Benotti and Blackburn (2021) also stress the importance of conversational grounding in deep learning based systems and on the ability to recover from mistakes as an important aspect for meaningful collaboration.

More recently, Jokinen (2024) notes the need for both symbol and conversational grounding of utterances by LLM-based dialogue systems to real-world events to avoid incorrect information. Anikina et al. (2025) surveys grounding along the dimensions of modalities, scope of grounded knowledge and the type of grounding (symbol/static vs conversational/dynamic). They focus on symbol grounding, and the dynamic nature of knowledge grounding, where the external knowledge changes during the conversation, rather than the interactive process that we focus on. Tolzin and Janson (2025) give a conceptual overview of the mechanisms (embodiment, social features, mental model, joint action, knowledge base) for building shared knowledge and how they are interrelated in human-agent interaction, reinforcing the interconnected nature of symbol and conversational grounding.

These works focus on the methods to perform symbol grounding while emphasising the potential for improved systems by employing conversational grounding. They also point to the lack of studies on, and therefore the need for, rigorous evaluation of grounding. Our work reviews strategies of conversational grounding in LLMs with a special focus on the evaluation of the grounding process in TOD.

3 Theoretical Background

Seminal work in philosophy and psycholinguistics explore the concept of grounding in human communication. In philosophy, Stalnaker (2002) describes the *common ground* as the set of propositions that participants in a conversation mutually accept, i.e. what they take for granted as background information. A proposition is part of the common ground if all participants believe it to be true and believe that all other participants also believe it to be true (Stalnaker, 1999). In psycholinguistics, Clark and Brennan (1991) describe human communication as a *joint activity* that requires collaboration between the participants. For successful communication, they need to coordinate on the content (the common ground) and the process of communication (updating the common ground). This collaborative process is called *grounding*.

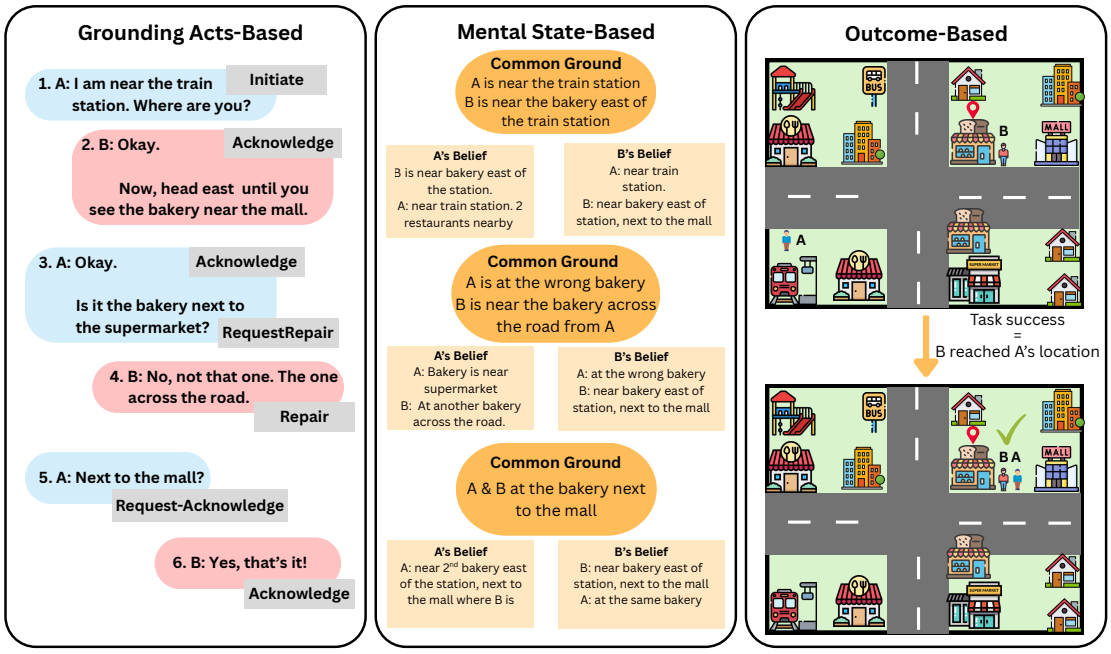
The Contribution Model Clark and Schaefer (1989) describe two phases in the grounding process: (i) the presentation phase, where interlocutor A presents an utterance for interlocutor B to consider; and (ii) the acceptance phase, in which B produces evidence of grounding. Together, these two phases make a *contribution*. The acceptance phase can cover several turns but it is necessarily concluded by a *positive* evidence of understanding, which can be verbal acknowledgments, a relevant next turn and gestures like a nod or unbroken eye gaze. However, *negative* evidences can also occur during this phase, as signs of mishearing and misunderstanding that should be collaboratively repaired in the upcoming turns. The grounding and its evidences change with the purpose and the medium of communication, while keeping the collaborative effort minimal to build contributions and reach mutual acceptance. For example, how you communicate differs when you are talking face-to-face or writing an email and whether you are giving out critical information or having a casual chat.

The contribution model defines a **grounding criterion**, which acts as a threshold, based on which, the decision to perform grounding can be made. Its formal definition is: "*The contributor and his/her partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for current purposes.*" In other words, if the grounding criterion is low, there is enough understanding and hence low need to perform grounding. Although it varies widely depending on the goal, the participants and the medium of communication, formalizing the grounding criterion helps in building and evaluating dialogue systems.

Seminal work proposed to formalize the grounding criterion and to perform grounding using either special dialogue acts called *grounding acts* (Traum, 1999) or decision making under uncertainty (Horvitz and Paek, 2000). In these two pieces of work and in general, the notion of grounding criterion directly connects to predicting the *utility* of performing grounding, i.e., if the utility of grounding is low then the grounding criterion is low. When the listener has comprehended the speaker's utterance, there is low utility in performing grounding and vice versa. From the speaker perspective, the utility of performing grounding depends on the speaker's belief that the listener has understood the utterance. Connecting back to Clark's contribution model, when the listener comprehends enough and the speaker believes that the listener has comprehended the utterance, the information is added to the common ground and a contribution is made.

A last important aspect is that performing grounding comes with some costs (time spent, cognitive effort, etc.). Hence, there is a trade-off to be estimated between maximizing the mutual understanding and minimizing the costs of the grounding process. For example, correcting a mistake as early as possible by providing strong evidence of grounding would have less temporal costs than correcting it later in the conversation. As a consequence, good grounding behaviour should take into consideration the utility *and* costs of executing the grounding action for the task at hand and whether it brings it closer to making a contribution.

Interactive Alignment Model Unlike the intentional joint activity in the contribution model, Pickering and Garrod (2004) views grounding as part of an automatic process introduced in the *Interactive Alignment Model* (IAM). Interlocutors align their mental states by aligning automatically at lower levels of representation. Alignment at the phonetic and phonological levels (using similar sounds) percolate upwards to the lexical (using the same words) and syntactic (same grammatical structure) levels, leading to the aligning of the mental states. The sources of alignment can be automatic like preconceived beliefs about one's interlocutor or through imitation of words, syntax or prosody (also known as *entrainment* (Kejriwal et al., 2026)). It can also result from feedback and agreement between interlocutors, similar to the pro-



(a) (b) (c)

Figure 2: An example dialogue demonstrating the three strategies for modelling conversational grounding.

Grounding Acts	Description
INITIATE	Initial utterance that proposes information to be grounded
CONTINUE	Continuation of the previous act from the same speaker
ACKNOWLEDGE	Acknowledgement of the presented information to show understanding
REPAIR	Correction of a previous utterance or addition of omitted material
REQREPAIR	Request a correction explicitly
REQACK	Request an acknowledgement from the other interlocutor
CANCEL	Abandons the grounding in the current discourse unit

Table 1: Grounding acts proposed by Traum and Allen (1992).

cess described by Clark and Brennan (1991). The information that is automatically available to the interlocutors as a result of their alignment at various levels is called the *implicit common ground*.

These theoretical frameworks offer complementary views on how interlocutors achieve mutual understanding. The rest of the paper considers Clark's contribution model as the foundational grounding theory, as it emphasizes the strategic, evidence-based and collaborative nature of conversations.

4 Evaluation of Conversational Grounding

Conversational grounding theory has been operationalised in various ways to make the grounding process observable and evaluable. This section distinguishes three main strategies: (*i*) using grounding acts, (*ii*) modelling the mental state of the participants to construct the common ground, and (*iii*) designing collaborative tasks. Grounding acts provide observable evidence of grounding, which influences the hidden mental states of participants, eventually leading to a positive or negative outcome of the task. For each strategy, we provide an overview of the implementations in the literature, and discuss the corresponding approaches to evaluate conversational grounding. Figure 2 illustrates the strategies.

4.1 Grounding Acts-Based Evaluation

Grounding Acts (GAs) (Traum and Allen, 1992) are special dialogue acts (Austin, 1962) that indicate the speaker's attempt to establish mutual understanding, as listed in Table 1. GAs were introduced as a simplification for the contribution model during on-the-fly processing of dialogues, with INITIATE indicating the presentation phase and ACKNOWLEDGE the acceptance phase. During a conversation (e.g. see Figure 2a), grounding involves *a priori* (i.e. based on dialogue history) deciding whether or not to use a GA and which GA

to use before generating the upcoming utterance. From the evaluation perspective, many works use the presence of GAs as an *a posteriori* signal (i.e. based on the full conversation) of the grounding process. Although recent studies use Clark's definition of grounding, they do not often follow the contribution model, which result in them choosing to define their own GAs or focus on smaller subsets. Table 3 in Appendix A lists the GAs used in recent papers.

Typology Among the grounding acts proposed by Traum and Allen (1992) (Table 1), ACKNOWLEDGE presents positive evidence of grounding. Many recent studies include explicit acknowledgements as an indication of grounding since it is easy to detect (Mohapatra et al., 2024a,b; Shaikh et al., 2024, 2025; Jokinen et al., 2024; Schneider et al., 2024). Another common GA is the REQUESTREPAIR, which is mostly found as clarification questions in dialogue (Shaikh et al., 2024; Jokinen et al., 2024; Schneider et al., 2024), followed by REPAIRS which clarifies the misunderstanding. These represent negative evidence seen as an indication of misalignment (Sarkar et al., 2025) or as friction that helps the system reflect and ground instead of assuming the user's needs (İnan et al., 2025).

Other evidence of grounding include REPEATBACK and USE (Roque and Traum, 2008) which demonstrate understanding by repeating or using the presented information in the next turn (Mohapatra et al., 2024a,b). However, these are not used widely in literature, probably due to the ambiguous nature of these acts—a repetition might mean understanding or that the user requires more clarity on the content. New acts like OVERRESPONSE have been introduced owing to the behaviour of LLMs that produce long and verbose responses (Shaikh et al., 2025).

Evaluation The grounding capability of LLMs is evaluated by measuring the GAs produced in LLM responses and by the LLM's ability to classify the GAs present in a turn. Turn-level GA classification is often implemented using prompting or fine-tuning (Mohapatra et al., 2024a,b; Shaikh et al., 2024, 2025; Jokinen et al., 2024; Schneider et al., 2024). Reference-based metrics such as precision, recall, F1 and accuracy are used. These works interpret grounding to be satisfactory the higher the accuracy of GA classification is. Current work shows limited performance, attributing it to the imbalance in GA categories. This highlights how important GAs like REPAIR are often underrepresented in current datasets (Mohapatra et al., 2024a). Zero-shot and few-shot prompting show that LLMs accurately classify GAs like ACKNOWLEDGE but tend to misclassify similar GAs (e.g. CLARIFICATION vs FOLLOWUP) (Shaikh et al., 2024; Schneider et al., 2024) or more implicit GAs (USE) (Jokinen et al., 2024; Mohapatra et al., 2024a). The low performance of LLMs in detecting REQUESTREPAIRS correlates negatively with the task success (Sarkar et al., 2025).

Another way of evaluating grounding in LLMs is by looking at how and when GAs are used. LLMs show very low inter-annotator agreement with humans with regards to the number of GAs used and its position in the conversation (Shaikh et al., 2024). The perplexity of a response using a correct and incorrect GA in a given context has been used to test whether the LLM uses the right GA in the right context. Lower perplexity for the correct response indicates the model is able to ground (Mohapatra et al., 2024b). Although perplexity is lower for larger models, the question remains whether it actually improves the model's understanding. By examining the distance between paraphrased utterances in the embedding space, it was found that models are unable to differentiate utterances when only the key information is changed (e.g "*on the yellow chair*" changed to "*on the yellow table*"). This indicates that models rely on lexical content rather than pragmatics (Mohapatra et al., 2024b).

4.2 Mental State-Based Evaluation

Following the theory of mind (Baron-Cohen, 1991), several works have attempted to model the mental states of the interlocutors in order to estimate the common ground as the intersection of their beliefs as shown in Figure 2b.

Existing Methods One approach is the *dialogue state tracking*, in which the cumulative belief state of the system about the user's goals (user's beliefs) is tracked at each utterance. The dialogue state is the first order belief of the system, what the system believes the user wants based on the user utterances (Jacqmin et al., 2022). The dialogue state was initially designed to handle uncertainty from noisy speech recognition (Williams and Young, 2007) and different methods for belief state tracking and evaluation have been widely studied (Williams et al., 2013; Henderson et al., 2014a,b; Kim et al., 2016a,b, 2019; Gunasekara

et al., 2020; Yoshino et al., 2024; Kottur and Moon, 2023). In recent years, the focus shifted towards zero-shot and few-shot prompting using LLMs to predict the dialogue state (Feng et al., 2023; Wang et al., 2024). The dialogue state guides the model on the next action to be taken.

Other works take belief modelling a step further by estimating both the first and second-order belief state of the participants (Qiu et al., 2024). The second order belief is what the interlocutor believes that the other participant believes that she/he believes. The goal is to reach common ground by reducing the difference in the first and second order beliefs through appropriate next turns.

A more fine-grained approach to modelling first-order belief state is based on evidences. Events are extracted from the utterances and the level of belief that participants have in them are predicted. The beliefs are then added to the common ground based on evidence from the conversation (Markowska et al., 2023; Khebour et al., 2024). These methods introduce challenges around event and belief extraction, evidence collection and propagation of errors into common ground estimation.

Evaluation The methods described above evaluate grounding by measuring how close the predicted beliefs are to the *optimal common ground* or a minimal amount of shared knowledge required to understand the conversation and complete the task. Dialogue state tracking is mainly evaluated using the Joint Goal Accuracy (JGA) (Jacqmin et al., 2022). JGA is a reference-based cumulative metric which evaluates the system's ability to keep track of the modifications of the user requests across turns. Instruction-tuned open source LLMs show improvement over state-of-the-art baselines on JGA (Feng et al., 2023; Wang et al., 2024). Other methods evaluate the predicted belief states by computing the F1 score against gold reference at turn-level (Qiu et al., 2024; Markowska et al., 2023). The performance of the belief module directly impacts the next step, whether it is generating the response (Qiu et al., 2024) or predicting the events in the common ground (Markowska et al., 2023).

Evidence-based methods for modelling the mental state employ similarity-based metrics. Markowska et al. (2023) proposes a similarity score called EMBERT that calculates the cosine similarity between embeddings of the extracted events and the gold reference. The Sorensen-Dice coefficient is used to compare the extracted propositions that were promoted to common ground through evidence with the ground truth, which improves with multimodal cues (Khebour et al., 2024). These studies show that there is a large scope for improvement in the systems used for event generation and assigning the right level of evidence, which impacts the prediction of the common ground.

These metrics rely on turns produced with a gold conversation history and a reference common ground, but in reality, the system may not follow the same conversation path. There will be multiple ways to contribute a piece of information to the conversation. Therefore, the scope of the common ground required to achieve the task varies based on how the conversation evolves.

4.3 Outcome-Based Evaluation

Identifying the optimal common ground and defining the grounding criterion is a challenging task. Hence, conversational grounding has been studied by using task success as a proxy to achieving common ground (See Figure 2c). Special tasks, known as *collaborative tasks*, have been designed to help evaluate grounding at the conversation level (Schlangen, 2019). Table 2 shows some collaborative tasks with the associated datasets.

Types of Collaborative Tasks *Alignment tasks* involve identifying the common element in a given situation through discussion leading to agreement on their final choice. They encourage the use of grounding actions such as clarifications, acknowledgement and referring expressions in a controlled setting to reach the goal faster. An example of an alignment task is **Mutual Friend** (He et al., 2017), where two participants with a private list of friends and information about them (e.g. name, school, etc.) converse to find their mutual friend.

Navigation tasks have the end goal of reaching a given location through dialogue. They help evaluate grounding as a consequence of the actions performed through strategically planned conversations—describing surroundings, correcting misunderstandings etc. An example is **MeetUp!** (Ilinykh et al., 2019), an interactive visual dialogue game where two players move in a visual environment of a grid of rooms, with the goal of meeting in the same room. They can navigate to different rooms and describe them to coordinate to meet in the same room.

While alignment and navigation tasks have par-

Dataset	Task	Modality	Automatic metrics used
Mutual Friend (He et al., 2017)	Alignment	Text	Success rate
Spot The Difference (Lopes et al., 2018)	Alignment	Text (audio transcript)	Task success
One Common (Udagawa and Aizawa, 2019)	Alignment	Text, image	Success rate
Dynamic One Common (Udagawa and Aizawa, 2021)	Alignment	Text, dynamic image	Success rate, length of success timestep
MeetUp! (Ilinykh et al., 2019)	Navigation	Text, image	Task success
Talk The Walk (de Vries et al., 2018)	Navigation	Text, image	Localisation accuracy
CaSiNo (Chawla et al., 2021)	Negotiation	Text	Points based on priority of items
Deal or No Deal (Lewis et al., 2017)	Negotiation	Text	Agreement rate, Pareto optimal

Table 2: Selected collaborative dialogue datasets that aid in studying conversational grounding at a global level. We have not included situated datasets like SIMMC here since the related papers focus on turn level tasks like coreference resolution, multimodal DST etc. and not on the task success.

ticipants working towards the same goal, *negotiation tasks* involve participants with adversarial goals trying to find a middle ground. They enable evaluation of grounding in more complex settings, since it involves strategic interactions that require understanding of preferences and constraints to maximise the success. An example is **CaSiNo** (Chawla et al., 2021), which involves two campsite neighbours with individual preferences negotiating for food, water and firewood.

We also mention here another category of tasks called *situated tasks*. These tasks require participants to interact with their environment to achieve a goal. These datasets are multimodal and hence, rich in both symbol and conversational grounding, allowing us to study the interaction between them. The **SIMMC** (Situated and Interactive MultiModal Conversations) family of datasets (Moon et al., 2020; Kottur et al., 2021; Wu et al., 2023) is a task-oriented dataset in the shopping domain, in which the user and assistant are in an evolving shared virtual reality context.

Evaluation Collaborative tasks focus on global evaluation of grounding with a proxy metric. The collaborative tasks shown in Table 2 mainly focus on task-specific success as the main indication that the participants have reached common ground. In alignment and navigation tasks, the task success depends on whether the participants choose the same entity (He et al., 2017; Udagawa and Aizawa, 2019, 2021) or reach the same location (Ilinykh et al., 2019; de Vries et al., 2018) at the end of the game. However, task success can be an ambiguous signal for grounding—participants may succeed due to a lucky guess despite very low mutual understanding.

In a negotiation task like CaSiNo (Chawla et al., 2021), the primary metric evaluating the conversation is the number of points obtained by the negotiator based on the priority of the items obtained. Depending on the task, other metrics like agreed % and pareto optimal (Lewis et al., 2017; Qiu et al., 2024) are used. These tasks also often include subjective human evaluation which rates the conversation on dimensions like satisfaction, fluency, correctness, cooperation and human-likeness (He et al., 2017; Chawla et al., 2021). However, these tasks do not enable us to measure how grounding is evolving throughout the conversation. While task success means that enough grounding is achieved, it does not take into account the time taken to complete the task at hand. Task completion without efficiency will not fare well in real world systems. We need to be able to measure whether the system is capable of taking corrective measures in order to reach grounding faster and hence successfully complete the task. For example, a repair would mean more turns, but once a repair is made, the goal can be reached efficiently. In contrast, not giving enough time for grounding actions may result in a more costly mistake later, which may result in an unsuccessful task but with fewer turns.

5 Avenues for Research

In this section, we describe future research directions regarding the use of conversational grounding in LLMs, its evaluation in TOD systems and its correlation with other evaluation dimensions.

5.1 Conversational Grounding in LLMs

The first direction is to encourage LLMs to ground better. In the context of GAs, one attempt is mitigation prompting, which prompts the LLM with explicit instructions to use clarifications and acknowledgements when necessary (Shaikh et al., 2024). Although the frequency of GAs in LLM responses increases, the agreement with how humans use GAs still remains low. Another method involves fine-tuning by rewarding the model for identifying the correct response and penalizing for selecting the incorrect response in a context where grounding is required (Mohapatra et al., 2024b;

Acikgoz et al., 2026). The inability of LLMs to use GAs correctly is attributed to preference training methods which specifically tune the models to follow instructions by assuming common ground (Shaikh et al., 2025).

Despite these efforts, LLMs still need to improve in identifying when and what GA to use in order to maximise the utility of performing grounding while lowering task-related costs in the long run. Future work should explore incorporating strategies for grounding in the various training stages of LLMs, anchored to Clark's theory of grounding. With many closed models in circulation, there is also a need for methods to handle grounding in black-box LLMs. An emphasis also needs to be put on the interconnected nature of symbol and conversational grounding. Improvements in linking external entities can reduce the need for repairs while effective clarifications can help reach symbol grounding more accurately, helping to develop better systems that can collaborate to achieve goals faster and adapt to new tasks as well (Lemon, 2022).

5.2 Grounding Evaluation

When developing methods to improve grounding, it is important to be able to evaluate their progress. GA detection relies on turn-level ground truths and accuracy scores to evaluate the presence of grounding. It does not take into account the progress in mutual understanding over time nor does it attempt to explain why a given act was used in a given turn and not another, which is an important aspect to address in evaluation. When modelling common ground through the mental states of participants, we should keep in mind that dialogue is dynamic, and different grounding behaviours can interact, which can influence the common ground that is established at the end of the conversation. Rather than relying on reference-based metrics, evaluation should attempt to measure how far the conversation might have strayed from the optimal common ground (i.e. the minimum shared knowledge required to achieve the goal) and how much each turn brings it closer to the common ground. It should also focus on the system's ability to bring the conversation back on track, to repair and recover from mistakes, and not just whether they make a mistake or not (Benotti and Blackburn, 2021).

Metrics that are model-agnostic and can be evaluated directly on the conversation are inevitable. An interesting approach can be to analyse internal representations of conversations to extract latent features that may act as evidence that correlate to grounding. Learning these representations could help detect the presence and level of grounding from the conversation itself.

5.3 Correlation of Grounding with Other Factors

Defining and evaluating a good conversation is not trivial. In TOD, task success rate, user satisfaction (Walker et al., 1997) and its objective counterpart, interaction quality (Schmitt et al., 2011) are some of the metrics used to evaluate the conversation. For better insights into the system performance, it would be interesting to study how the level of grounding impacts these metrics and find the right balance between them. A system that constantly clarifies may achieve perfect task success but can frustrate users, reducing satisfaction. On the other hand, a system that does not ground at all might come across as very efficient to the user while the task solution does not meet the user requirements.

Correlation of grounding with other dimensions like trust, skills of the participants and the dialogue strategies, to name a few, can also be studied. Whether the system is able to build a representation of the speaker in addition to that of the conversation can be interesting to explore. For example, a skilled participant can adapt to the listener's expertise to decide what information to introduce and how much it needs to be grounded for the concerned task. This could improve the perceived trust of the participants in each other to solve the task.

6 Conclusion

As the community moves toward more capable agents, conversational grounding remains a crucial aspect in the path forward. Even though LLMs generate very fluent responses, a key factor for effective human-machine collaboration is the ability to pause, reflect and ground the conversation, rather than making assumptions about the context.

In this paper, we have highlighted various methods for modelling conversational grounding in recent systems, how they are used in the evaluation of grounding and the challenges they face. Progress in grounding capabilities requires evaluation frameworks that can test different grounding scenarios (recovery from mistakes, identifying and clarifying ambiguity, etc.), measure the balance between the level of grounding and other evaluation dimensions and account for the interaction between symbol

and conversational grounding. To that effect, we have suggested a few promising research avenues for the community to consider, that motivate more rigorous, robust and model-agnostic metrics for conversational grounding.

Limitations

While this paper is a survey on the evaluation methods of conversational grounding in LLMs, it is not exhaustive. Our aim is to provide a general understanding of conversational grounding, how it is commonly modelled and associated evaluation techniques. Several works focusing on only a single aspect of grounding (e.g. acknowledgements or repairs) have not been included. Interested readers can refer to Di Maro (2021) for a comprehensive overview of works studying each of the original grounding acts. Similarly, much work has been done on dialogue state tracking, including several surveys (Jacqmin et al., 2022) which explore this topic in depth.

Acknowledgement

This work was funded by the ANRT CIFRE contract number 2024/1692.

References

Emre Can Acikgoz, Jinoh Oh, Jie Hao, Joo Hyuk Jeon, Heng Ji, Dilek Hakkani-Tur, Gokhan Tur, Xiang Li, Chengyuan Ma, and Xing Fan. 2026. SpeakRL: Synergizing reasoning, speaking, and acting in language models with reinforcement learning. In *Proceedings of the 16th International Workshop on Spoken Dialogue System Technology*, pages 312–325, Trento, Italy. Association for Computational Linguistics.

Tatiana Anikina, Alina Leippert, and Simon Ostermann. 2025. Building common ground in dialogue: A survey. In *Proceedings of the 2nd LUHME Workshop*, pages 3–28, Bologna, Italy. LUHME.

John L. Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford [Eng.].

Simon Baron-Cohen. 1991. Precursors to a theory of mind: Understanding attention in others. *Natural theories of mind: Evolution, development and simulation of everyday mindreading*, 1(233-251):1.

Luciana Benotti and Patrick Blackburn. 2021. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

Herbert H. Clark and Susan Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *Preprint*, arXiv:1807.03367.

Maria Di Maro. 2021. Computational grounding: an overview of common ground applications in conversational agents. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):133–156.

Michelle Elizabeth, Morgan Veyret, Miguel Couceiro, Ondrej Dusek, and Lina M. Rojas Barahona. 2025. Exploring ReAct prompting for task-oriented dialogue: Insights and shortcomings. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 143–153, Bilbao, Spain. Association for Computational Linguistics.

Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. 2023. Towards LLM-driven dialogue state tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 739–755, Singapore. Association for Computational Linguistics.

Charisse Foo, Pin Sym Foong, Camille Nadal, Natasha Ureyang, Thant Naylin, and Gerald Choon Huat Koh. 2025. The benefits and risks of llms for facilitating medical decision-making among laypersons. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, DIS '25, page 3173–3191, New York, NY, USA. Association for Computing Machinery.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, Dilek Hakkani-Tür, Jinchao Li, Qi Zhu, Lingxiao Luo, Lars Liden, Kaili Huang, Shahin Shayandeh, Runze Liang, Baolin Peng, and 20 others. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *Preprint*, arXiv:2011.06486.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329.

Eric Horvitz and Tim Paek. 2000. Grounding criterion: Toward a formal theory of grounding. Technical Report MSR-TR-2000-40.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet up! a corpus of joint activity dialogues in a visual environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.

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.

Kristiina Jokinen. 2024. The need for grounding in LLM-based dialogue systems. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 45–52, Torino, Italia. ELRA and ICCL.

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. Towards harnessing large language models for comprehension of conversational grounding. *CoRR*, abs/2406.01749.

Jay Kejriwal, Štefan Beňuš, and Lina M. Rojas-Barahona. 2026. Entrainment detection using DNN. *Computer Speech & Language*, 99:101930.

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia. ELRA and ICCL.

Seokhwan Kim, Luis D'Haro, Rafael Banchs, Jason Williams, and Matthew Henderson. 2016a. The fourth dialog state tracking challenge.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E. Banchs, Jason D. Williams, Matthew Henderson, and Koichiro Yoshino. 2016b. The fifth dialog state tracking challenge. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517.

Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, Minlie Huang, Luis Lastras, Jonathan K. Kummerfeld, Walter S. Lasecki, Chiori Hori, Anoop Cherian, Tim K. Marks, Abhinav Rastogi, Xiaoxue Zang, and 2 others. 2019. The eighth dialog system technology challenge. *Preprint*, arXiv:1911.06394.

Satwik Kottur and Seungwhan Moon. 2023. Overview of situated and interactive multimodal conversations (SIMMC) 2.1 track at DSTC 11. In *Proceedings of the Eleventh Dialog System Technology Challenge*, pages 235–241, Prague, Czech Republic. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Topics in Cognitive Science*, 10(2):389–408.

Oliver Lemon. 2022. Conversational grounding in emergent communication – data and divergence. In *Emergent Communication Workshop at ICLR 2022*.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark. Association for Computational Linguistics.

José Lopes, Nils Hemmingsson, and Oliver Åstrand. 2018. The spot the difference corpus: a multi-modal corpus of spontaneous task oriented spoken interactions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024a. Conversational grounding: Annotation and analysis of grounding acts and grounding units. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3967–3977, Torino, Italia. ELRA and ICCL.

Biswesh Mohapatra, Manav Nitin Kapadnis, Laurent Romary, and Justine Cassell. 2024b. Evaluating the effectiveness of large language models in establishing conversational grounding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9767–9781, Miami, Florida, USA. Association for Computational Linguistics.

Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Shuwen Qiu, Mingdian Liu, Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024. MindDial: Enhancing conversational agents with theory-of-mind for common ground alignment and negotiation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 746–759, Kyoto, Japan. Association for Computational Linguistics.

Antonio Roque and David Traum. 2008. Degrees of grounding based on evidence of understanding. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 54–63, Columbus, Ohio. Association for Computational Linguistics.

Rupak Sarkar, Neha Srikanth, Taylor Pellegrin, Rachel Rudinger, Claire Bonial, and Philip Resnik. 2025. Understanding common ground misalignment in goal-oriented dialog: A case-study with Ubuntu chat logs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3200–3215, Vienna, Austria. Association for Computational Linguistics.

David Schlangen. 2019. Grounded agreement games: Emphasizing conversational grounding in visual dialogue settings. *CoRR*.

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, pages 173–184, Portland, Oregon. Association for Computational Linguistics.

Phillip Schneider, Nektarios Machner, Kristiina Jokinen, and Florian Matthes. 2024. Bridging information gaps in dialogues with grounded exchanges using knowledge graphs. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 110–120, Kyoto, Japan. Association for Computational Linguistics.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Omar Shaikh, Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2025. Navigating rifts in human-LLM grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20832–20847, Vienna, Austria. Association for Computational Linguistics.

Robert Stalnaker. 1999. *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press UK, Oxford, GB.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line active reward learning for policy optimisation in spoken dialogue systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441, Berlin, Germany. Association for Computational Linguistics.

Antonia Tolzin and Andreas Janson. 2025. Uncovering the mechanisms of common ground in human–agent interaction: review and future directions for conversational agent research. *Internet Research*.

David R Traum. 1999. Computational models of grounding in collaborative systems. In *Psychological Models of Communication in Collaborative Systems-Papers from the AAAI Fall Symposium*, pages 124–131.

David R. Traum and James F. Allen. 1992. A "speech acts" approach to grounding in conversation. In *ICSLP*.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Takuma Udagawa and Akiko Aizawa. 2021. Maintaining common ground in dynamic environments. *Transactions of the Association for Computational Linguistics*, 9:995–1011.

Marilyn A. Walker, 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.

Wenxuan Wang, Shi Juluan, Zixuan Ling, Yuk-Kit Chan, Chaozheng Wang, Cheryl Lee, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2025a. Learning to ask: When LLM agents meet unclear instruction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21773–21784, Suzhou, China. Association for Computational Linguistics.

Xiaoye Wang, Nicole Xi Zhang, Hongyu He, Trang Nguyen, Kun-Hsing Yu, Hao Deng, Cynthia Brandt, Danielle S. Bitterman, Ling Pan, Ching-Yu Cheng, James Zou, and Dianbo Liu. 2025b. Safety challenges of ai in medicine in the era of large language models. *Preprint*, arXiv:2409.18968.

Xingguang Wang, Xuxin Cheng, Juntong Song, Tong Zhang, and Cheng Niu. 2024. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8724–8741, Bangkok, Thailand. Association for Computational Linguistics.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

Jason D. Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Comput. Speech Lang.*, 21(2):393–422.

Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. 2023. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6273–6291, Toronto, Canada. Association for Computational Linguistics.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. 2025. Towards Visual Grounding: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–20.

Koichiro Yoshino, Yun-Nung Chen, Paul Crook, Satwik Kottur, Jinchao Li, Behnam Hedayatnia, Seungwhan Moon, Zhengcong Fei, Zekang Li, Jinchao Zhang, Yang Feng, Jie Zhou, Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papangelis, Karthik Gopalakrishnan, Dilek Hakkani-Tur, Babak Damavandi, and 9 others. 2024. Overview of the tenth dialog system technology challenge: Dstc10. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:765–778.

Mert İnan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. 2025. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *Preprint*, arXiv:2501.17348.

A Appendix

Table 3 shows the different grounding acts used in recent literature, further categorized into one of the seminal set of grounding acts by Traum and Allen (1992) and additional finegrained acts by Roque and Traum (2008).

Grounding Acts	Term used in recent literature	Description	Associated Papers
INITIATE⋆		Initial utterance that proposes information to be grounded (the presentation phase)	Mohapatra et al. (2024a,b)
	Restart	Reset a conversation and reinitialise the same information to ground	Shaikh et al. (2025)
CONTINUE⋆		Continuation of the previous act from the same speaker	Mohapatra et al. (2024a,b)
ACKNOWLEDGE⋆		Acknowledgement of the presented information to show understanding	Mohapatra et al. (2024a,a)
	Acknowledgement	Explicit use of utterances like Okay, Yeah etc.	Shaikh et al. (2024, 2025)
REPAIR⋆		Correction of a previous utterance	Mohapatra et al. (2024a,b); Shaikh et al. (2025); Sarkar et al. (2025)
	Reformulation	Repeats or restates the query in other words due to a failure to ground	Shaikh et al. (2025)
REQREPAIR⋆		Request a correction explicitly	Mohapatra et al. (2024a,b); Sarkar et al. (2025)
	Clarification	Further clarification asked by interlocutor	Shaikh et al. (2024, 2025); Schneider et al. (2024); Jokinen et al. (2024)
REQACK⋆		Request an acknowledgement from the other interlocutor	Mohapatra et al. (2024a,b)
CANCEL⋆		Abandons the grounding in the current discourse unit	Mohapatra et al. (2024a,b)
SUBMIT†		Presenting a piece of information that has not yet been mentioned and has no assumed values	
REPEATBACK†		Repeat the same information back as a form of explicit confirmation	Mohapatra et al. (2024a,b)
RESUBMIT†		Information already presented is corrected (shows negative evidence)	
MOVE ON†		Move to the next step in the conversation	
	Next turn	Utter a relevant next turn to move forward in the conversation	Shaikh et al. (2025)
USE†		Use the information provided	Mohapatra et al. (2024a,b)
	Follow-up	Ask for elaboration on a prior utterance	Shaikh et al. (2024, 2025)
LACK OF RESPONSE†		Neither participant gives a response for a given length of time	
Other	Overresponse	Response contains more information than what was asked	Shaikh et al. (2025)

Table 3: Grounding acts categorization based on Traum and Allen (1992) (⋆) and Roque and Traum (2008) (†).