Identifying Repair Opportunities in Child-Caregiver Interactions

Abhishek Agrawal, Benoit Favre, and **Abdellah Fourtassi**
Aix Marseille Univ, CNRS, LIS, Marseille, France
Correspondence: abhishek-amit.agrawal@univ-amu.fr

Abstract

Identifying and repairing breakdowns in communication is critical for maintaining and developing common ground between interlocutors. In the context of child-caregiver interaction (the focus of the current study), parents' repair has an important role both for scaffolding early communication and for providing a learning signal. Developmental studies have typically focused on linguistic markers of parents' repair and repair initiation (e.g., clarification request); here, we analyze such instances relative to overall repair *opportunities*, allowing us to quantify the extent to which parents seize repair opportunities. We analyzed a corpus of $N = 15$ child-caregiver dyads, where children were 7 to 11 years old. Using manual annotation, we identified the repair opportunities presented by the child and found that the caregivers initiated repair in only a small subset of the cases where a repair opportunity presented itself. We further tasked several large language models (LLMs) to test their capacity to recognize repair opportunities in children's utterances and found their performance to be lacking compared to human annotators. This study provides an initial exploration that is valuable both for developmental studies and for researchers aiming to improve child–machine interaction for applications such as personalized education.

1 Introduction

For an effective, intelligible, and fluent conversation, a key competency that the interlocutors must possess is the ability to successfully coordinate and negotiate their shared beliefs, knowledge, and assumptions (Clark and Schaefer, 1989; Clark, 1996; Stalnaker, 1978). This ability — also known as *conversational grounding* — allows the interlocutor to interpret an utterance accurately based on their shared knowledge with the speaker of the utterance and the dialog history, thereby letting the interlocutor respond in a coherent and effective manner. It helps the interlocutors resolve any ambiguity and clear up misunderstandings that occur during a conversation (Fried et al., 2023).

Interlocutors in a conversation start out with some shared belief space or *common ground* from shared culture, a social group, or previous interaction (Baker et al., 1999; Clark, 1996). The common ground is then further developed throughout the conversation by contributions from all the participants in the conversation (Clark and Brennan, 1991). For grounding any information, the interlocutors need to provide implicit or explicit evidence that information has been well communicated and understood. This evidence can take the form of acknowledgments (e.g., backchannels), initiation of the relevant next turn, by showing continued attention (e.g., through eye gaze), by issuing a clarification request, among other signals (Clark and Brennan, 1991; Clark and Krych, 2004; Clark and Schaefer, 1989).

Identifying and repairing breakdowns in early communication

Typically, an interlocutor has a communicative intent in mind that they need to get across to the other interlocutor. To illustrate, suppose Jane has an intent I in mind and tries to communicate it to Jack. If I has something to do with a situated object — as, for instance, is often the case when talking to a young child — then one of the actions Jane can take is to simply point to the object to indicate her intent, or by looking at the target, inviting gaze following (e.g., Frank et al., 2009). However, if I is not situated — e.g., an abstract idea or a displaced target — something that becomes more and more prevalent as children develop, then Jane and Jack need to ground I in their mutual understanding by more sophisticated means. To this end, Jane continually monitors Jack for signs of understanding of her intent. If Jack shows signs of misunderstanding, then she can step in and repair the misunderstand-

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.

ing. The objective of this exercise is to ensure that both Jack and Jane share their understanding of Jane's intent I.

Indeed, one of the primary mechanisms for maintaining common ground is identifying and repairing breakdowns in communication (Clark and Krych, 2004; Purver et al., 2018; Benotti and Blackburn, 2021; Dingemanse et al., 2015; Fusaroli et al., 2017; Schegloff, 1992). However, we know little about how this mechanism plays out in child development, especially in child-caregiver interactions. This is a significant gap given that many proposals suggest a role for caregivers' communicative feedback on children's production, especially when these productions are ill-coordinated or poorly constructed and potentially helping in furthering language use refinement (Nikolaus and Fourtassi, 2023; Clark, 2018, 2020). While there is a wealth of studies focusing on caregivers' role in guiding infants' understanding when in a situated context and the target is visually accessible (e.g., review in Çetinçelik et al., 2021), there is hardly any study quantifying this phenomenon when the target is abstract or not visually available to interlocutors.

A notable difficulty here, especially when analyzing spontaneous conversations (e.g., CHILDES, MacWhinney, 2000), is that the intent to be grounded (i.e., I) is not always apparent to a third party, namely the researcher, making the analysis fully dependent on the caregiver's reaction to what the child said (e.g., whether the caregiver asked for clarification). While a focus on the caregiver reactions allows for an estimate of actual repair initiation, this estimate can be misleading because it does not account for all *repair opportunities*, some of which may have been missed or ignored by caregivers.

To address the difficulty of identifying the caregiver's intent in fully unstructured settings, here we resort to using a weakly structured word-guessing game that allows us to maintain a (relatively) naturalistic conversational style while also providing access to the ground truth intent I (i.e., the word to be guessed). While this context — where the caregiver is making a child guess a word — is not fully naturalistic, it is meant to approximate the instance when the caregiver and child work collaboratively to ground a complex intent or idea (e.g., why limiting screen time is important) in shared understanding, only here this intent is operationalized, for simplicity, as a simple word that needs to be guessed.

The goals of the current study

Using this setup, a first goal of the current study is to quantify caregivers' actual repair relative to repair opportunities, as follows. First, we characterize all children's questions (e.g., "Does this object fly?") in terms of being well or ill-coordinated, thanks to our access to the caregiver's intent and the exchange history. In particular, the subset of children's questions that are ill-coordinated (e.g., asking "Does this object fly?" when it was already established that the object cannot take flight) provides the set of what we call **repair opportunities**. Second, we characterized instances of caregivers' actual repair (e.g., the caregiver reminding the child that their question is not valid given what has been discussed so far).

A second goal was to study the extent to which Large Language Models (LLMs) can recognize repair opportunities in children's utterances; a fundamental task these models need to solve in order to be able to provide effective repair and help in children's learning (e.g., in an personalized educational setting), in a similar way that caregivers' repair help children learn (e.g., Clark, 2020). We examine the capabilities of current LLMs to identify whether the child's question is valid or not given the previous conversational context and the word picked by the caregiver. Figures 1 and 2 demonstrate this experimental setup.

Our annotations and our code to run all the models locally are publicly available at `https://github.com/abhishek-agrawal94/common-ground`.

2 Methods

2.1 Data

We make use of the CHICA corpus (Goumri et al., 2024) which consists of recordings of child-caregiver face-to-face interactions in French. There are 15 dyads across three age groups in middle childhood (5 recordings per group) where the age of the child is around 7, 9 and 11 years old. The interlocutors take turns in picking a word and having the other interlocutor try to guess the word correctly by asking various questions about it. For further details on the data collection we refer the reader to the original paper.

2.2 Manual Annotation

After masking all the personal identifiers of the interlocutors in the data, we manually annotated all

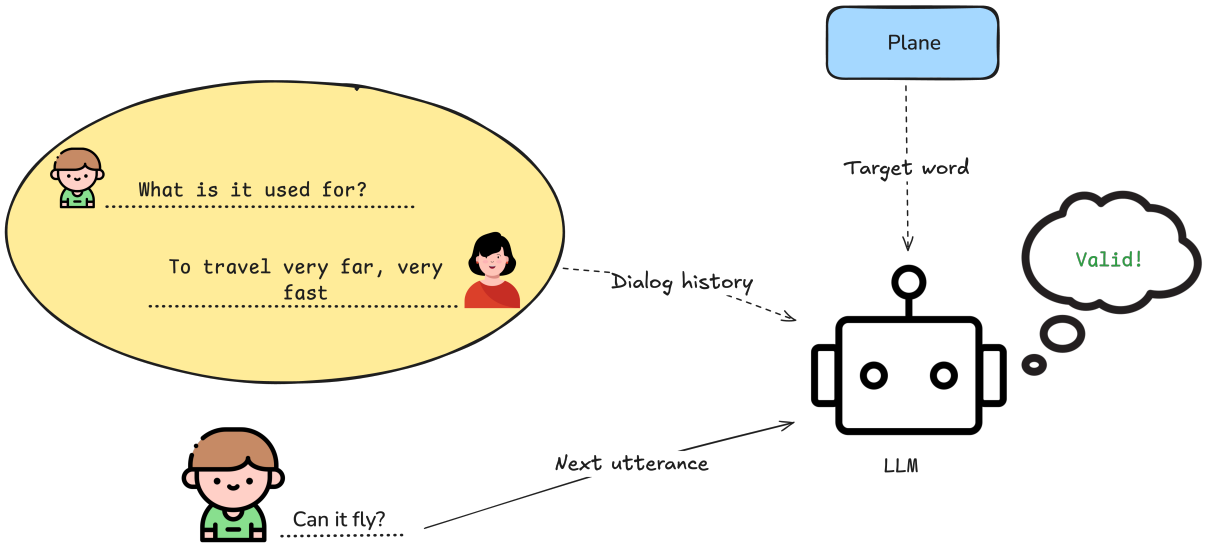

Figure 1: Example of a valid question asked by the child.

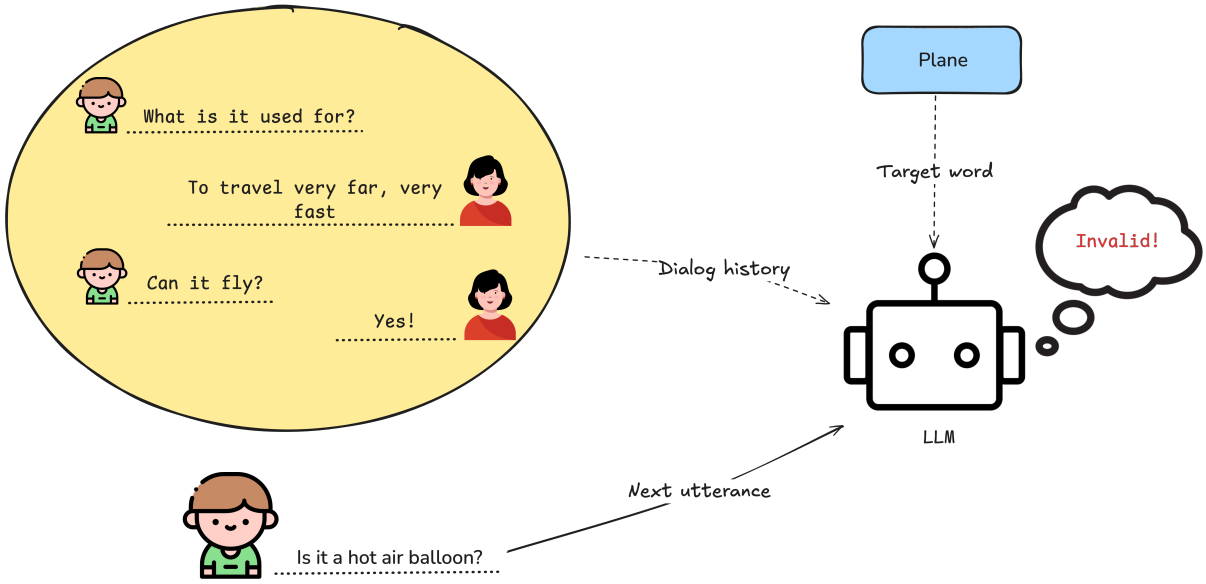

Figure 2: Example of an invalid question asked by the child leading to a possible repair opportunity. Here, the question is invalid because a hot air balloon can neither travel very far nor is it very fast.

the questions asked by the child as either *"valid"* or *"invalid"* based on the previous dialog history and the word being guessed (which is known to the caregiver but not to the child). We considered only the transcript of the conversation while annotating the data. A child's question was marked as invalid if the question directly contradicted some information or a fact that was established by the parent and the child in the past dialog turns. Questions were also marked as invalid if they were repetitions of the same questions that were previously asked by the child. Two authors annotated approximately 25% of the data separately and obtained a Cohen's Kappa score of $\kappa = 0.75$. The first author annotated the rest of the data, leading to a total of $N = 739$ questions across the entirety of the 15 recordings. In addition to these repair opportunities, we also annotated whether the caregiver initiated a repair.[1]

2.3 LLMs' testing

We tested a variety of models on our task of identifying breakdowns in child-caregiver interactions. Our selection of models are from amongst the most widely used set of models which have generally shown good performance across several tasks on

[1]These repairs are all other-repairs; we didn't annotate for self-repairs.

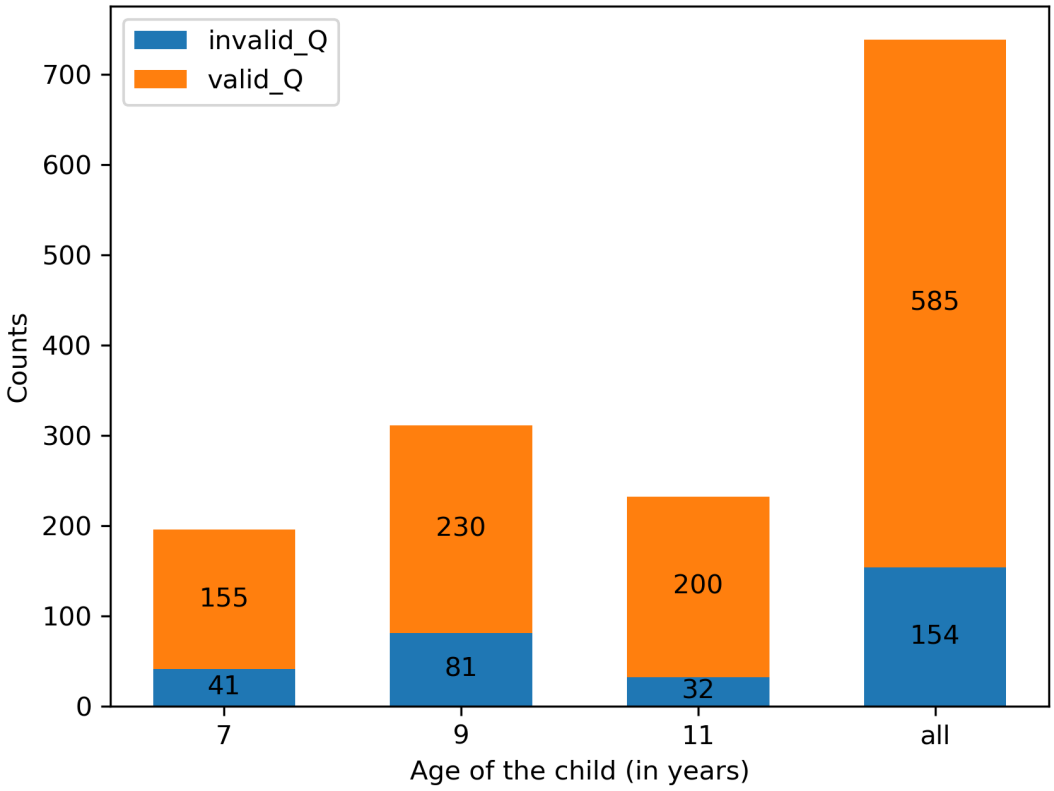

Figure 3: Distribution of valid and invalid questions asked by the child across all age groups.

various benchmarks and leader-boards (for e.g., the Chatbot Arena (Chiang et al., 2024)). All our models (except for GPT-4o) were downloaded from Ollama[2] and run locally on our system in inference mode (no fine-tuning). All the models downloaded from Ollama are 4-bit quantized versions by default. The models we tested in our study are as follows:

- Llama-3.1 8B[3]
- Llama-3.2 3B[4]
- Gemma-2 9B[5]
- Phi-3 14B[6]
- Mistral 7B[7]
- Mistral-nemo 12B[8]
- GPT-4o[9]

We used a few-shot prompting strategy to elicit from the LLMs whether a question posed by the child to the caregiver is valid or not based on all the previous relevant dialog history until that point. The prompt templates can be found in the Appendix A. We tested the LLMs both on the original French data as well as the English translation[10] to see if the language of communication affected the performance of the models.

3 Results and Analyses

3.1 Caregiver repairs vs. repair opportunities

First, we show the results of manual annotation. Figure 3 shows the distribution of children's breakdowns in our manually annotated data. 585 questions in total were "valid" and 154 questions were

[2] https://ollama.com/
[3] https://huggingface.co/meta-llama/Llama-3.1-8B
[4] https://huggingface.co/meta-llama/Llama-3.2-3B
[5] https://huggingface.co/google/gemma-2-9b
[6] https://huggingface.co/microsoft/Phi-3-medium-128k-instruct
[7] https://mistral.ai/news/announcing-mistral-7b
[8] https://mistral.ai/news/mistral-nemo
[9] https://openai.com/index/hello-gpt-4o/
[10] Obtained through Google Translate and manual correction

Language Model	Balanced accuracy	
	English	French
Llama-3.1	0.60	0.59
Llama-3.2	0.57	0.57
Gemma-2	0.69	0.69
Mistral	0.61	0.62
Mistral-nemo	0.58	0.58
Phi-3	0.62	0.59
GPT-4o	**0.75**	**0.76**
Human score	0.84	

Table 1: Balanced accuracy scores for few-shot prompting strategy.

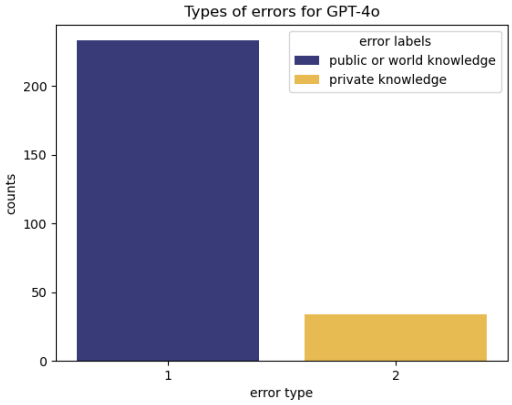

Figure 4: Error analysis for GPT-4o model.

"invalid", showing insufficient grounding of prior information in around 26% of the time. This number varied across age groups (26% in the younger age group, 35% in the middle, 16% in the older group), but these numbers do not reflect a systematic developmental change.

The set of invalid questions represent what we call repair opportunities. We found that caregiver initiated $N = 59$ repairs, 95% of which followed invalid questions. Thus, out of a total of 154 repair opportunities, caregiver instantiated repair in about 36% of the time. Thus, while caregiver repair is not rare, it addresses only a minority of repair opportunities.

3.2 Can LLMs detect repair opportunities?

Table 1 shows the balanced accuracy scores for all the models when identifying whether a question by the child is valid or not. As seen in the table, the score of all models (except GPT-4o) are generally low and barely perform above chance, showcasing the difficulty of the task. This was the case both when using the original version in French and when using the English translation, showing that the reasons the models find the task difficult is not due to the use of French (as one may suspect, given that the models are trained primarily on English data scraped off the internet). While GPT-4o (and to some extent Gemma-2) shows a much better accuracy (around 0.75 in the case of GPT-4o), it is still lower than accuracy based on human inter-annotation agreement (Cohen's Kappa score of $\kappa = 0.75$ translates into an accuracy of 0.84).

Error analysis We analyzed the errors that these models make in their predictions. An interesting distinction to consider in the context of grounding is when the information to be grounded is a) common world knowledge that even strangers can have access to, and b) when this information is, instead, more dependent on the interlocutors sharing previous experiences that a third party may not have access to. We gave real examples of both cases from our data, shown in boxes 1 (Example 1) and 2 (Example 2). In example 1, which illustrates errors regarding common world knowledge, the caregiver has a target word "book", they establish with the child that the object is rectangular, after which the child poses the question, "Does that mean a square?". The child's question can be easily classified as "invalid" by a third party. In example 2, which illustrates annotation difficulty due to private knowledge, the caregiver's target word is "cactus" and the child asks if they have the object available at home. The child's question relates to private knowledge (what the dyad has at home) and, therefore, makes the question less straightforward to categorize.

Given that the models do not have access to the interlocutors' private common ground, it could be understandable if most errors fall in this category. However, this was not the case. In fact, the overwhelming majority of the errors (Figure 4 for the best performing model) are related to common world knowledge (around 87%) and only a small subset concerns private knowledge shared by the interlocutor (around 13%), indicating ample room for improvement in these models regarding common world knowledge reasoning.

Example 1

Target word: **A book**

Caregiver: *Usually it is rectangular.*
Child: *What does a rectangle mean?*
Caregiver: *It's like a rectangle. And there are words inside.*
Child: *What is a rectangle again?*
Caregiver: *A rectangle is like that.*
Child: *Like that?*
Caregiver: *Like your presentation folder.*
Child: *Ok.*
Child: *Does that mean a square?* → [**Invalid**]
Caregiver: *No, a rectangle.*

— ChiCa corpus, ID_3.csv

Example 2

Target word: **A cactus**

Child: *Is it green?*
Caregiver: *Yeah, it's green.*
Child: *Is it a bit hard?*
Caregiver: *Yes, except when it rots, it becomes very soft.*
Caregiver: *It's hard and above all it has a special characteristic.*
Child: *Does it smell good?*
Caregiver: *No, not particularly.*
Child: *Is it hard?*
Caregiver: *Yes.*
Child: *Is that... I don't know.*
Caregiver: *I don't know, ask if we have any or does it grow, I don't know.*
Child: *Do we have any?* → [**Valid**]
Caregiver: *We have a whole one... In one of the planters, there are some very small ones.*

— ChiCa corpus, ID_2.csv

Does caregiver's repair indicate severity of communicative breakdown? So far, we tested the models on their ability to detect all repair opportunity and we found low to moderate performance. However, we know that caregivers initiate repair in only a minority of cases (as we saw above). One possibility is that caregivers initiate repair only for the subset of invalid questions that are more severe and risk to seriously disrupt the grounding process, in which case, we would expect the models to find it easier to classify those more apparent cases. As reported above, caregivers initiated a total of $N = 59$ repairs, 56 of which followed invalid questions and 3 followed valid questions. To create a balanced testing data, we randomly sampled 56 valid questions and 3 invalid ones. Table 2 shows the results when we restricted our analysis to this subset. The accuracies are very similar to when the models were tested on the larger dataset; suggesting that the subset of repairs that caregivers initiate do not necessarily target more obvious cases of communicative breakdown, at least from the perspective of the LLMs we tested.

Language Model	Accuracy
Llama-3.1	0.60
Llama-3.2	0.55
Gemma-2	0.67
Mistral	0.63
Mistral-nemo	0.56
Phi-3	0.64
GPT-4o	0.75

Table 2: Accuracy scores for repair initiating questions.

4 Conclusions

This study offers a first exploration into the identification of repair opportunities in child–caregiver interactions. The main finding is that caregivers address only a small portion—approximately one third—of the potential repair opportunities that arise during conversation.

We also evaluated the ability of several large language models (LLMs) to identify repair opportunities in children's utterances. Compared to human annotators, the models showed limited performance, underscoring the complexity of the task. Among the models tested, the larger, closed-weight model GPT-4o outperformed the smaller, open-weight models, consistent with prior findings on grounding-related tasks (Hakimov et al., 2025; Mohapatra et al., 2024b). Error analysis revealed that GPT-4o's failures often stem from a lack of common world knowledge and/or limitations in reasoning over such knowledge.

As an initial exploration, this study comes with several limitations. Like any corpus-based analysis, as opposed to experimental approaches, it does not allow for the elicitation of specific phenomena and is constrained by what occurs naturally in the data. For example, although we annotated a relatively large number of data points (N=739), instances of

our target phenomenon—invalid questions, or repair opportunities—were limited to just 154 cases. This smaller sample size limits the strength of our conclusions. Accordingly, our main finding, that parents respond to only a subset of available repair opportunities, should be interpreted with caution and awaits confirmation in future, larger-scale studies.

A key limitation in our evaluation of LLMs' ability to identify repair opportunities is that the models merely "overheard" the conversation, rather than actively participating in it. As Madureira and Schlangen (2024) rightly argue, developing common ground requires active engagement in the interaction. However, an 'overhearing' paradigm does not undermine the outcome of our approach, since the task focused solely on assessing the validity of questions based on prior dialogue context, rather than simulating the grounding process itself, a step that we did not address here.

Finally, our study relied solely on transcripts. However, the development of common ground is inherently a multimodal phenomenon. Evidence of grounding and repair initiation often appears in visual cues such as head nods, shakes, frowns, or pointing gestures. In naturalistic, free-flowing conversation, the signals for detecting and addressing communication breakdowns are typically more subtle and multimodal than what was captured here. That said, this limitation is partially mitigated by the design of the game, which required interlocutors to verbalize their repair initiations—making them almost always identifiable in the transcripts. While this reduces concerns about the internal validity of our operationalization (by focusing on the transcript), it leaves open the broader question of ecological validity.

To conclude, our corpus analysis reveals that caregivers draw on only a limited subset of the potential repair opportunities that arise during interactions with children. Moreover, we show that several large language models (LLMs) underperform compared to humans in identifying these opportunities for repair in child–caregiver conversations. This highlights the need for further improvement, particularly in the context of applying LLMs to e-tutoring systems.

5 Related work

Recently, LLMs' limitations in conversational grounding abilities have drawn the eye of NLP researchers (Benotti and Blackburn, 2021; Chandu et al., 2021; Shaikh et al., 2024). For instance, studies show that LLMs struggle to understand when utterances are implicitly grounded (Jokinen et al., 2024) and that if conversational agents are augmented with theory-of-mind modeling, it improves their capacity to align with the speaker and helps in negotiating their common ground (Qiu et al., 2024). Cheng et al. (2024) explore the use of a multimodal transformer model to predict uncertainty in young children engaged in a counting task and finds a potential for improvement. The uncertainty of an interlocutor during a conversation is evidence of their understanding, which ties in directly to their shared common ground. Benotti and Blackburn (2021) raise an important concern that the way current LLMs and dialog models interact with humans can be misleading at times as they build false expectations of their common ground with their interlocutors. This, in turn, leads to a rise in misunderstandings, which can be frustrating for humans attempting to converse with dialogue models.

One of the avenues of studying grounding is to consider using LLMs to evaluate grounding acts in a conversation as proposed by Traum (Traum and Allen, 1992). LLMs tend to struggle to classify grounding acts in a conversation as well as generate them and their classification capabilities are directly linked to the number of parameters of the model and the size of it's pre-training data (Mohapatra et al., 2024a; Shaikh et al., 2024; Mohapatra et al., 2024b). Various forms of grounding have been studied by considering several grounded language tasks like reference games (for e.g., Kennington and Schlangen, 2015; Golland et al., 2010; Monroe et al., 2017) and goal-oriented dialog tasks (for e.g., Das et al., 2017; De Vries et al., 2017; Kim et al., 2019; Udagawa and Aizawa, 2019; Haber et al., 2019; Narayan-Chen et al., 2019) amongst other things (see Chandu et al. (2021) for a non-exhaustive list of tasks). However, the issue with most of these tasks and the phenomena that the researchers are trying to model with these tasks is not really conversational grounding and it is restricted to either referential grounding or grounding in terms of a particular modality (e.g., visual grounding) (Ilinykh et al., 2019; Hakimov et al., 2025; Jeknić et al., 2024; Chandu et al., 2021).

Acknowledgments

This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (France 2030), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). Furthermore, this study was also supported by the ANR MACOMIC (ANR-21-CE28-0005-01) grant. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011013886).

References

M. Baker, Tia G. B. Hansen, R. Joiner, and D. Traum. 1999. The role of grounding in collaborative learning tasks.

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. *arXiv preprint*. ArXiv:2106.02192 [cs].

Qi Cheng, Mert İnan, Rahma Mbarki, Grace Grmek, Theresa Choi, Yiming Sun, Kimele Persaud, Jenny Wang, and Malihe Alikhani. 2024. Learning Multimodal Cues of Children's Uncertainty. *arXiv preprint*. ArXiv:2410.14050.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Eve V. Clark. 2018. Conversation and Language Acquisition: A Pragmatic Approach. *Language Learning and Development*, 14(3):170–185.

Eve V. Clark. 2020. Conversational Repair and the Acquisition of Language. *Discourse Processes*, 57(5-6):441–459.

Herbert H. Clark. 1996. *Using Language*. 'Using' Linguistic Books. Cambridge University Press, Cambridge.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Perspectives on socially shared cognition.*, pages 127–149. American Psychological Association, Washington.

Herbert H. Clark and Meredyth A. Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1):62–81.

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

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. *arXiv preprint*. ArXiv:1611.08669 [cs].

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual Object Discovery through Multi-modal Dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475, Honolulu, HI. IEEE.

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):e0136100. Publisher: Public Library of Science.

Michael C Frank, Noah D Goodman, and Joshua B Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, 20(5):578–585.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. Pragmatics in Language Grounding: Phenomena, Tasks, and Modeling Approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.

Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten H. Christiansen, and Mark Dingemanse. 2017. Measures and mechanisms of common ground: backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In *Annual Meeting of the Cognitive Science Society*.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 410–419, Cambridge, MA. Association for Computational Linguistics.

Dhia Elhak Goumri, Abhishek Agrawal, Mitja Nikolaus, Hong Duc Thang Vu, Kübra Bodur, Elias Emmar, Cassandre Armand, Chiara Mazzocconi, Shreejata Gupta, Laurent Prévot, Benoit Favre, Leonor Becerra-Bonache, and Abdellah Fourtassi. 2024. CHICA: A Developmental Corpus of Child-Caregiver's Face-to-face vs. Video Call Conversations in Middle Childhood. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), pages 3153–3164, Torino, Italia. ELRA and ICCL.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Sherzod Hakimov, Yerkezhan Abdullayeva, Kushal Koshti, Antonia Schmidt, Yan Weiser, Anne Beyer, and David Schlangen. 2025. Using Game Play to Investigate Multimodal and Conversational Grounding in Large Multimodal Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5686–5718, Abu Dhabi, UAE. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. MeetUp! A Corpus of Joint Activity Dialogues in a Visual Environment. *arXiv preprint*. ArXiv:1907.05084 [cs].

Isidora Jeknić, David Schlangen, and Alexander Koller. 2024. A Dialogue Game for Eliciting Balanced Collaboration. *arXiv preprint*. ArXiv:2406.08202 [cs].

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. Towards Harnessing Large Language Models for Comprehension of Conversational Grounding. *arXiv preprint*. ArXiv:2406.01749 [cs].

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. CoDraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513, Florence, Italy. Association for Computational Linguistics.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed.* The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US. Pages: xi, 366.

Brielen Madureira and David Schlangen. 2024. It Couldn't Help But Overhear: On the Limits of Modelling Meta-Communicative Grounding Acts with Supervised Learning. *arXiv preprint*. ArXiv:2405.01139 [cs].

Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. 2024a. Conversational Grounding: Annotation and Analysis of Grounding Acts and Grounding Units. In *LREC-COLING 2024 - The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Turin, Italy.

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.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

Mitja Nikolaus and Abdellah Fourtassi. 2023. Communicative feedback in language acquisition. *New Ideas in Psychology*, 68:100985.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational Models of Miscommunication Phenomena. *Topics in Cognitive Science*, 10(2):425–451.

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.

Emanuel A. Schegloff. 1992. Repair After Next Turn: The Last Structurally Provided Defense of Intersubjectivity in Conversation. *American Journal of Sociology*, 97(5):1295–1345. Publisher: The University of Chicago Press.

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.

Robert C. Stalnaker. 1978. Assertion. Brill. Section: Pragmatics.

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 AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Melis Çetinçelik, Caroline F. Rowland, and Tineke M. Snijders. 2021. Do the eyes have it? a systematic review on the role of eye gaze in infant language development. *Frontiers in Psychology*, Volume 11 - 2020.

A Prompt Template

The prompt templates for the English and French transcripts are shown in Fig 5 and 6 respectively.

Two people are playing a word guessing game where player 1 picks a word and
player 2 doesn't know this word. Player 2 needs to ask questions to player 1
to guess the word correctly. Given the dialog history in terms of the turns taken
by player 1 and player 2 and the word picked by player 1, you need to decide whether
the next question asked or statement made by player 2 or the object mentioned by player 2 is valid or
not based on the dialog history until that point. You need to give a boolean binary response (True or False) whether
the question is valid or not in JSON format. Use the following template: {valid: ""}.

Here are some examples to help you out.

Example 1: Word picked by player 1: A balloon.
Dialog history: player 2 turn: Is it a living being? player 1 turn: No. player 2 turn: Is it an object? player 1 turn: Yes.
Next question: Can you play with it?
{valid: True}

Example 2: Word picked by player 1: A cat.
Dialog history: player 2 turn: Is it a living being? player 1 turn: Yes. player 2 turn: Can it be a pet? player 1 turn: Yes.
Next question: a cat?
{valid: True}

Example 3: Word picked by player 1: A car.
Dialog history: player 2 turn: Is it a living being? player 1 turn: No.
Next question: is it an insect?
{valid: False}

End of examples.

Word picked by player 1: <TARGET_WORD>
Dialog history: <DIALOG_HISTORY>
Next question: <QUESTION>

Figure 5: Prompt template with English examples and transcript.

Two people are playing a word guessing game in the French language where player 1 picks a word and player 2 doesn't know this word. Player 2 needs to ask questions to player 1 to guess the word correctly. Given the dialog history in terms of the turns taken by player 1 and player 2 and the word picked by player 1, you need to decide whether the next question asked or statement made by player 2 or the object mentioned by player 2 is valid or not based on the dialog history until that point. You need to give a boolean binary response (True or False) whether the question is valid or not in JSON format. Use the following template: {valid: ""}.

Here are some examples to help you out.

Example 1: Word picked by player 1: Un ballon.
Dialog history: player 2 turn: Est-ce que ça un être vivant? player 1 turn: Non.
player 2 turn: Est-ce que ça un objet? player 1 turn: Oui.
Next question: Peux-tu jouer avec ça?
{valid: True}

Example 2: Word picked by player 1: Un chat.
Dialog history: player 2 turn: Est-ce que ça un être vivant? player 1 turn: Oui.
player 2 turn: Est-ce que ça peut être un animal de compagnie? player 1 turn: Oui.
Next question: un chat?
{valid: True}

Example 3: Word picked by player 1: Une voiture.
Dialog history: player 2 turn: Est-ce que ça un être vivant? player 1 turn: Non.
Next question: Est-ce que ça un insecte?
{valid: False}

End of examples.

Word picked by player 1: <TARGET_WORD>
Dialog history: <DIALOG_HISTORY>
Next question: <QUESTION>

Figure 6: Prompt template with French examples and transcript.