Misalignments in Common Ground as a Bridge Between Pragmatic Theory and LLM Evaluation

Judith Sieker and Sina Zarrieß
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{j.sieker;sina.zarriess}@uni-bielefeld.de

Abstract

In this position paper, we argue that misalignments in common ground are not marginal failures of communication, but central diagnostic moments for pragmatic competence, and should therefore play a key role in the evaluation of Large Language Models (LLMs). Evaluating how models respond to such instances of mismatched or incomplete understanding moves beyond surface fluency and correctness, targeting pragmatic competence at a deeper, interactional level. At the same time, misalignments provide controlled settings for testing linguistic theories of common ground, repair, or accommodation – areas that are often difficult to investigate in human communication. We argue that this dual role makes misalignments a natural bridge between pragmatic theory and LLM evaluation.

1 Introduction

At the heart of successful communication lies the concept of common ground – the shared knowledge, beliefs, and assumptions that speakers rely on to understand one another. This shared foundation allows conversations to flow smoothly, with speakers and listeners working together to maintain mutual understanding (Clark, 1996). Yet, in everyday communication, achieving and maintaining common ground is rarely flawless. Time constraints and the need to prioritize information often lead speakers to make assumptions about their audience's knowledge (Beaver, 1999), which can result in misalignments in common ground (Elder and Beaver, 2022). Such misalignments are not rare exceptions but a routine part of communication. For example, imagine the scenario where a local gives directions to a tourist. The local might say, "Turn right after where Miller's bakery used to be," presupposing that the listener knows this bakery and where it was located. A tourist unfamiliar with the area lacks this shared knowledge, resulting in a misalignment in common ground. These moments signal where shared assumptions fail and must be actively negotiated and resolved in order for the interaction to progress, requiring pragmatic strategies such as repair or accommodation. Without resolving misalignments, conversations risk stalling, misunderstandings, or the spread of false information. In this position paper, we argue that misalignments in common ground offer a valuable lens for studying communicative competence, not despite, but because they introduce challenges requiring pragmatic reasoning. They provide a window into whether a language user can detect when shared understanding has broken down and take appropriate steps to restore it.

As Large Language Models (LLMs) are now widely deployed in everyday applications, it becomes increasingly important to assess their ability not just to produce fluent and informative text, but also their capacity to engage in context-sensitive dialogue that is robust to misunderstanding. It remains largely an open question to what extent they can detect and respond to situations where assumptions misalign, expectations are violated, or clarification is needed. By examining LLM behavior in such cases, we gain insight into how they manage the very conditions that define real-world communication. At the same time, modeling and testing how humans resolve such misalignments remains challenging, creating an opportunity for LLMs to serve as controlled testbeds for investigating theories of common ground and communicative grounding. In other words, LLMs can serve not only as objects of evaluation, but also as tools for probing pragmatic theories under controlled conditions. We therefore argue that misalignments in common ground provide a natural bridge between formal pragmatics and computational evaluation, allowing insights from linguistic theory and model behavior to inform one another.

Third Workshop on the Bridges and Gaps between Formal and Computational Linguistics, pages 74–81

2 Misalignments in Common Ground

The concept of common ground has been highly influential in the study of dialogue (Bender and Lascarides, 2019) and plays a central role in many areas of pragmatics, including reference, presupposition, implicature, speech acts, language conventions, and fiction (Geurts, 2024). However, common ground does not arise automatically in conversation. Rather, speakers and listeners work collaboratively to establish and maintain it, a process known as communicative grounding (Larsson, 2018; Chandu et al., 2021). Communicative grounding involves ensuring that participants accurately perceive, understand, and accept each other's utterances (Chandu et al., 2021). To accomplish this, speakers commonly engage in dialogue acts such as asking for clarification or posing follow-up questions (Shaikh et al., 2024). While this process can be straightforward, it often requires addressing and resolving misalignments in the common ground (Larsson, 2018).

Misalignments in common ground refer to instances of mismatched or incomplete understanding between participants in a conversation. Unlike simple misunderstandings, misalignments specifically involve a gap or mismatch in what interlocutors assume to be shared knowledge. They can be overt – recognized and possibly acknowledged – or covert – unrecognized and therefore more difficult to resolve. Furthermore, misalignments in common ground can take different forms, including inconsistencies in what interlocutors believe to be shared knowledge, or gaps that reflect missing or incomplete information (Elder and Beaver, 2022). For example, a speaker referencing "last week's meeting" may misassume that the listener was present. Or a local giving directions may presuppose knowledge the tourist lacks. In both cases, successful communication depends on resolving the misalignment, either implicitly through accommodation, where the listener silently adjusts their knowledge to align with the speaker's presuppositions (von Fintel, 2008; Beaver et al., 2024), or explicitly through repair, such as asking for clarification or correcting the assumption (Clift, 2014; Birkner et al., 2020). Such everyday cases highlight that misalignments are a routine consequence of the fact that no speaker has exhaustive knowledge of the world; assumptions frequently fail to align perfectly with a listener's perspective.

A central question is what conditions influence whether conversation participants choose to accommodate or repair misalignments. Existing research has explored, for example, the social incentives behind repair, highlighting it as an explicit acknowledgment of communication problems (Robinson, 2014), or as a socially motivated strategy (Bernstein, 2016; Elder and Beaver, 2022). However, modeling when and how speakers respond to misalignments remains challenging. While theoretical frameworks provide first insights (Elder and Beaver, 2022), empirical and cognitive research suggests that speakers do not always behave fully collaboratively, complicating attempts to model their behavior in a unified way. For instance, speakers may act egocentrically (Horton and Keysar, 1996; Keysar, 2007), overestimate alignment with their interlocutor (Keysar and Henly, 2002), or have only partial awareness of the common ground (Brown-Schmidt, 2012). Compounding this, not every apparent misalignment reflects a genuine breakdown: hearers sometimes interpret unfamiliar references as informative rather than presupposition-failing, effectively accommodating the new information rather than triggering explicit repair (Heller et al., 2012). As a result, even for human communication, misalignments are difficult to investigate systematically.

3 LLMs and the Challenge of Grounding

To address this challenge, LLMs may offer a useful testbed: they allow misalignments to be studied under controlled conditions and can thus serve as tools for probing theories of accommodation and repair. More broadly, recent work has argued that LLMs can function as testbeds for evaluating linguistic and psychological hypotheses under controlled conditions (Contreras Kallens et al., 2023; Demszky et al., 2023; Millière, 2024). For instance, their scale enables testing across a wide range of contexts, and their intervenability allows specific aspects of the input or model behavior to be manipulated. At the same time, LLMs face key limitations: Handling pragmatic phenomena depends on world knowledge and social norms, and LLMs differ from humans in important ways, for example in their training conditions and lack of social interaction (Hu, 2023; Mahowald et al., 2023; Millière, 2024). This highlights the need for such considerations when using LLMs to study linguistic phenomena and theories, including those related to (misalignments in) common ground.

Beyond their role as experimental tools, LLMs are themselves participants in real-world communicative settings, making it essential to understand their pragmatic behavior in practice. These models have demonstrated strong performance on tasks like summarization and question answering (Millière, 2024), and user-facing tools such as ChatGPT (Achiam et al., 2023) have enabled widespread public experimentation. At the same time, LLMs can remain sensitive to surface-level input variation and may generalize inconsistently across linguistic contexts, pointing to limitations in their language use (Chang and Bergen, 2024). This has motivated a growing body of linguistically-informed research probing the underlying knowledge and behavior of LLMs. Much of this work focuses on syntactic abilities, where LLMs often perform near-human-level (Hu et al., 2020; Schuster and Linzen, 2022; Mahowald et al., 2023). Other work has extended this line of inquiry to semantic and pragmatic phenomena. Here, earlier studies frequently identified clear shortcomings: models faced limitations when interpreting sarcasm (Hu et al., 2023), understanding implicature and presuppositions (Cong, 2022; Ruis et al., 2022; Fried et al., 2023; Sieker and Zarrieß, 2023), handling implicit causality (Zarrieß et al., 2022; Sieker et al., 2023), or responding appropriately to theory-of-mind reasoning (Sap et al., 2022; Shapira et al., 2023; Trott et al., 2023; Gandhi et al., 2024)), suggesting weaknesses at the discourse level of language production (Chang and Bergen, 2024; Contreras Kallens et al., 2023). More recent studies, however, paint a more nuanced picture: especially larger and more recent LLMs can exhibit improved performance on certain pragmatic phenomena, including, for instance, selected aspects of implicature (Cong, 2024; Askari et al., 2025), implicit causality (Kankowski et al., 2025a,b), elicitur (Pan and Kehler, 2025), or politeness (Andersson and McIntyre, 2025).

Crucially, despite these advances, systematic evaluations of how LLMs engage in conversational grounding remain scarce. Most work has focused on assessing and improving their static knowledge (Fierro et al., 2024; Xu et al., 2024), largely overlooking the fact that effective communication requires aligning knowledge dynamically with an interlocutor (Larsson, 2018). In particular, systematic studies of communicative grounding in instruction-tuned LLMs like GPT-4 are only just beginning to emerge. Earlier work has qualitatively examined grounding failures in pretrained models, highlighting their limitations in establishing and maintaining shared understanding (Fried et al., 2023). Benotti and Blackburn (2021), for example, find that systems such as BlenderBot frequently fail to repair misunderstandings or establish shared meaning, suggesting a lack of mechanisms for collaborative grounding. Findings by Lee et al. (2024) and Shaikh et al. (2024) further support the observation that LLMs tend to assume shared understanding rather than actively negotiating it, likely reflecting their optimization for prompt following rather than for the nuanced and collaborative coordination that grounding requires. This pattern is consistent with research on LLM sycophancy, which shows that models often prioritize user approval over factual accuracy, aligning responses with user beliefs even when these are false (Perez et al., 2023; Rrv et al., 2024). More broadly, Chandu et al. (2021) observe that NLP research frequently overlooks key aspects of grounding emphasized in Cognitive Science, such as coordination and mutual understanding, revealing a conceptual gap that may underlie these persistent shortcomings.

Although these studies demonstrate initial progress in operationalizing grounding, relevant tasks and systematic benchmarks for evaluating common ground are still limited. One particularly overlooked aspect is how LLMs resolve misalignments in common ground – a crucial component of effective communication that, to date, has not been systematically investigated. A notable exception is Sarkar et al. (2025), who study misalignments in goal-oriented Ubuntu chat logs and show that LLMs particularly struggle to detect implicit ones.

Handling such cases requires more than grammaticality or lexical choice: it involves interpreting intentions, recognizing when communication has broken down, and deciding whether to accommodate or repair. This, in turn, demands consideration of the broader communicative context – something that has largely been absent from existing probing studies. This gap is especially concerning given the increasing use of LLMs in sensitive settings like healthcare or legal assistance, where pragmatic failures can have serious consequences. For instance, if LLMs fail to resolve misalignments in common ground but still respond fluently and confidently, they risk reinforcing user misconceptions (Fried et al., 2023). Investigating how LLMs manage common ground and repair misalignments is therefore not only a theoretical concern but also crucial for their responsible deployment.

4 Misalignments as a Lens for Evaluation

Misalignments in common ground are not marginal anomalies in communication; they are a natural, frequent, and revealing part of interaction. We argue that such misalignments should not be treated as noise or edge cases in LLM evaluation, but as core phenomena for assessing pragmatic competence. Furthermore, we argue that when investigating misalignments in common ground, we can assign LLMs a dual role. On the one hand, they are objects of evaluation: studying how LLMs respond to misalignments provides a direct probe of their pragmatic competence. On the other hand, they can serve as tools for investigation, as their controllability enables the construction of experimental settings in which assumptions about shared knowledge can be systematically varied – something that is difficult to achieve in studies of human interaction. In this way, misalignments offer a bridge between formal and computational approaches, linking theoretical accounts of common ground to empirical evaluation.

One way to instantiate such an evaluation paradigm is through linguistically well-defined phenomena. As an illustrative example, consider presuppositions. Presuppositions are implicit assumptions that speakers take for granted as shared knowledge (Stalnaker, 1973). They are essential for understanding how speakers rely on shared background knowledge, without making it explicit (Schwarz, 2019). For instance, the question "Has your sister finished reading the book I lent her?" conveys multiple presuppositions: that the listener has a sister, that the speaker lent her a book, and that she started reading it. One intriguing aspect of presuppositions, central to the study of misalignments in common ground, is *presupposition failure*, which occurs when the assumed background knowledge is actually false (Yablo, 2006). For instance, if the listener has no sister, is not aware of their sister borrowing the book, or if she has not actually started reading it, a presupposition fails, potentially causing a misalignment in common ground.

Recent work has shown that such false presuppositions can be used to probe LLMs' pragmatic competence in handling misalignments in common ground. For instance, Sieker et al. (2025) and Lachenmaier et al. (2025) systematically embedded false presuppositions in prompts, analyzing whether LLMs were able to reject the implied assumptions. With their evaluation framework, the authors show that false presuppositions are especially useful for the evaluation of misalignments in common ground: while the misalignment introduced by presuppositions is triggered implicitly, it requires an explicit intervention. Importantly, beyond demonstrating limitations in models' ability to handle such cases, the authors also show that linguistically controlled factors – such as the type of presupposition trigger or contextual framing – systematically affect model behavior. This allows model responses to inform not only assessments of pragmatic competence, but also questions central to linguistic theories of presupposition.

More broadly, linguistic phenomena such as presuppositions, implicatures (Grice, 1975), or elicitures (Cohen and Kehler, 2021), for example, provide theory-informed, minimal-pair-friendly conditions that can serve as evaluation paradigms to systematically probe misalignments in common ground. We encourage the use of such phenomena to develop evaluation frameworks that make misalignments experimentally controllable.

A practical way to operationalize this approach is to construct minimal pairs that systematically manipulate whether a misalignment is present. To illustrate this with a different phenomenon, consider implicature.

A minimal pair could consist of two contexts that are identical except for one piece of shared information. In context A (no misalignment), the prior discourse is neutral. A teacher states, "I graded the exams", after which the utterance "Some of the students passed the exam" licenses the standard implicature that not all students passed. In context B (misalignment), the prior discourse establishes that all students passed the exam. For instance, the teacher states, "Everyone did really well this time", after which the same utterance "Some of the students passed" creates a mismatch between the implicature and the shared contextual information.

In both contexts, a model is then presented with a neutral follow-up question, and its response can be scored on whether it ignores the tension and accommodates silently, or explicitly repairs it, for instance, by flagging the misalignment before answering. A key advantage of using LLMs in this paradigm is their controllability at multiple levels (Demszky et al., 2023). At the level of the prompt, linguistic factors can be systematically varied. For instance, scalar items ("some" vs. "many" vs. "most") can be substituted to test whether im-

plicature strength affects model behavior, or discourse contexts can be varied to examine how different settings affect the detection and resolution of misalignments. At the level of the model, (open-source) LLMs allow for intervention in training data, model size, or fine-tuning. In this way, LLMs can serve as a controlled testbed for investigating how implicatures are derived and interpreted and, more generally, how misalignments in common ground are detected and resolved.

5 Conclusion

In this position paper, we have argued that misalignments in common ground are not exceptions to effective communication but central to its structure, and that they create a concrete setting in which linguistic theory and LLM-based experimentation can productively inform one another. By examining how LLMs behave in situations where shared assumptions fail, we gain a twofold perspective: instances of misalignment provide theory-informed diagnostics of pragmatic competence, while LLMs offer controlled testbeds for investigating questions central to linguistic theory. We therefore invite the research community to place greater emphasis on the diagnostic moments of misalignments in common ground, and to develop tools and benchmarks that make communicative grounding and misalignment resolution a core part of model evaluation.

Limitations

There are limitations to this paper.

First, we focus on presuppositions and implicatures as illustrative examples because they are linguistically well-defined, amenable to minimal-pair designs, and supported by established experimental traditions. Misalignments in common ground, however, can arise through a much broader range of pragmatic phenomena. Future work should therefore investigate how the proposed framework extends to additional sources of misalignment.

Second, our example evaluation paradigm focuses on single-turn interactions, which allow for controlled, minimal-pair-based evaluation but cannot fully capture the dynamic, multi-turn negotiation of meaning that characterizes communicative grounding (Clark, 1996). Such paradigms nevertheless provide a useful starting point, offering a level of experimental control that is difficult to achieve in fully interactive settings. Exploring how controlled pragmatic evaluation can be extended to multi-turn interactions represents an important direction for future work. Recent work by Lachenmaier et al. (2026), for instance, demonstrates how linguistically grounded notions of repair can be operationalized in a multi-turn setting, pointing toward promising directions for such extensions.

Third, we do not address broader pragmatic capacities such as Theory of Mind (Holterman and van Deemter, 2023), which are closely related to common-ground reasoning but represent a distinct dimension of communicative competence. Understanding how such abilities interact with the detection and resolution of misalignments remains an important direction for future research.

Acknowledgements

We want to thank the anonymous reviewers for their very helpful comments and suggestions.

We acknowledge financial support from the project “SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Marta Andersson and Dan McIntyre. 2025. Can chatgpt recognize impoliteness? an exploratory study of the pragmatic awareness of a large language model. *Journal of Pragmatics*, 239:16–36.

Raha Askari, Sina Zarrieß, Özge Alacam, and Judith Sieker. 2025. Are BabyLMs deaf to Gricean maxims? a pragmatic evaluation of sample-efficient language models. In *Proceedings of the First BabyLM Workshop*, pages 52–65, Suzhou, China. Association for Computational Linguistics.

D Beaver. 1999. Presupposition accommodation: a plea for common sense. pages 21–44.

David I. Beaver, Bart Geurts, and Kristie Denlinger. 2024. Presupposition. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Emily M Bender and Alex Lascarides. 2019. Linguistic fundamentals for natural language processing II: 100 essentials from semantics and pragmatics. *Synth. Lect. Hum. Lang. Technol.*, 12(3):1–268.

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.

Katie A Bernstein. 2016. "misunderstanding" and (mis)interpretation as strategic tools in intercultural interactions between preschool children. *Appl. Linguist. Rev.*, 7(4):471–493.

Karin Birkner, Peter Auer, Angelika Bauer, and Helga Kotthoff. 2020. *Einführung in Die Konversationsanalyse*. de Gruyter Studium. De Gruyter, Berlin, Germany.

Sarah Brown-Schmidt. 2012. Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, 27(1):62–89.

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.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

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

Rebecca Clift. 2014. 4. conversation analysis. In *Pragmatics of Discourse*, pages 97–124. DE GRUYTER, Berlin, München, Boston.

Jonathan Cohen and Andrew Kehler. 2021. Conversational eliciture. *Philosophers' Imprint*, 21(12).

Yan Cong. 2022. Psycholinguistic diagnosis of language models' commonsense reasoning. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 17–22, Dublin, Ireland. Association for Computational Linguistics.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, and Morten H Christiansen. 2023. Large language models demonstrate the potential of statistical learning in language. *Cogn. Sci.*, 47(3):e13256.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Chi-He Elder and David Beaver. 2022. "we're running out of fuel!": When does miscommunication go unrepaired? *Intercult. Pragmat.*, 19(5):541–570.

Constanza Fierro, Ruchira Dhar, Filippos Stamatiou, Nicolas Garneau, and Anders Søgaard. 2024. Defining knowledge: Bridging epistemology and large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16096–16111, Miami, Florida, USA. Association for Computational Linguistics.

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.

Kanishk Gandhi, J.-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2024. Understanding social reasoning in language models with language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Bart Geurts. 2024. Common Ground in Pragmatics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Winter 2024 edition. Metaphysics Research Lab, Stanford University.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Daphna Heller, Kristen S. Gorman, and Michael K. Tanenhaus. 2012. To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science*, 4(2):290–305.

Bart Holterman and Kees van Deemter. 2023. Does chatgpt have theory of mind? *Preprint*, arXiv:2305.14020.

William S. Horton and Boaz Keysar. 1996. When do speakers take into account common ground? *Cognition*, 59(1):91–117.

Jennifer Hu. 2023. *Neural language models and human linguistic knowledge*. Ph.D. thesis, Massachusetts Institute of Technology.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment

of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Florian Kankowski, Torgrim Solstad, Sina Zarriess, and Oliver Bott. 2025a. Implicit causality-biases in humans and llms as a tool for benchmarking llm discourse capabilities. *Preprint*, arXiv:2501.12980.

Florian Kankowski, Torgrim Solstad, Sina Zarrieß, and Oliver Bott. 2025b. Instruction tuning modulates discourse biases in language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, pages 2818–2826.

Boaz Keysar. 2007. Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, 4(1).

Boaz Keysar and Anne S Henly. 2002. Speakers' overestimation of their effectiveness. *Psychological Science*, 13(3):207–212.

Clara Lachenmaier, Hannah Bultmann, and Sina Zarrieß. 2026. Talking to a know-it-all gpt or a second-guesser claude? how repair reveals unreliable multi-turn behavior in llms. *Preprint*, arXiv:2604.19245.

Clara Lachenmaier, Judith Sieker, and Sina Zarrieß. 2025. Can LLMs ground when they (don't) know: A study on direct and loaded political questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14956–14975, Vienna, Austria. Association for Computational Linguistics.

Staffan Larsson. 2018. Grounding as a side-effect of grounding. *Top. Cogn. Sci.*, 10(2):389–408.

Hyunji Lee, Se June Joo, Chaeeun Kim, Joel Jang, Doyoung Kim, Kyoung-Woon On, and Minjoon Seo. 2024. How well do large language models truly ground? 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 2437–2465, Mexico City, Mexico. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models.

Raphaël Millière. 2024. Language models as models of language. *Preprint*, arXiv:2408.07144.

Dingyi Pan and Andrew Kehler. 2025. Pragmatic competence in LLMs: The case of eliciture. In *Proceedings of the Society for Computation in Linguistics 2025*, pages 388–391, Eugene, Oregon. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Jeffrey D. Robinson. 2014. What "what?" 'tells us about how conversationalists manage intersubjectivity. *Research on Language and Social Interaction*, 47(2):109–129.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12717–12733, Bangkok, Thailand. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *arXiv [cs.CL]*.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. 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.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.

Florian Schwarz. 2019. *Presuppositions, Projection, and Accommodation*. Oxford University 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.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

Judith Sieker, Oliver Bott, Torgrim Solstad, and Sina Zarrieß. 2023. Beyond the bias: Unveiling the quality of implicit causality prompt continuations in language models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 206–220, Prague, Czechia. Association for Computational Linguistics.

Judith Sieker, Clara Lachenmaier, and Sina Zarrieß. 2025. LLMs struggle to reject false presuppositions when misinformation stakes are high. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

Judith Sieker and Sina Zarrieß. 2023. When your language model cannot Even do determiners right: Probing for anti-presuppositions and the maximize presupposition! principle. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 180–198, Singapore. Association for Computational Linguistics.

Robert Stalnaker. 1973. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Kai von Fintel. 2008. What is presupposition accommodation, again? *Philosophical Perspectives*, 22(1):137–170.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

S Yablo. 2006. Non-catastrophic presupposition failure.

Sina Zarrieß, Hannes Groener, Torgrim Solstad, and Oliver Bott. 2022. This isn't the bias you're looking for: Implicit causality, names and gender in German language models. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 129–134, Potsdam, Germany. KONVENS 2022 Organizers.