Modeling the Use-Mention Distinction in LLM-Generated Grounding Acts

Milena Belosevic
German Linguistics, Faculty of Linguistics and Literary Studies,
Bielefeld University
milena.belosevic@uni-bielefeld.de

Hendrik Buschmeier
Digital Linguistics Lab, Faculty of Linguistics and Literary Studies,
Bielefeld University
hbuschme@uni-bielefeld.de

Abstract

Given that large language models (LLMs) are systems that do not understand human language in a human-like way, LLM-generated grounding acts, such as explicit claims of understanding (e.g., "I understand"), can lead to overtrust in the capabilities of LLM chatbots, supporting their perception as human interlocutors (Shaikh et al., 2024). This paper argues for enriching these grounding acts with metalinguistic markers (e.g., scare quotes) that motivate users to perceive them as 'mentioned' and not as 'used' language (use–mention distinction; Sperber and Wilson, 1981). We illustrate how different types of meta-language can be enriched with (non)verbal metalinguistic units to mark LLM-generated grounding acts as mentioned language.

1 Introduction

Shared understanding is crucial for effective dialogues in human interactions and, arguably, interactions with artificial interlocutors. Therefore, a growing body of research deals with the role of common ground in interactions with LLMs (Jokinen et al., 2024; Mohapatra, 2023; Shaikh et al., 2024; Pilán et al., 2024). Defining common ground in the context of LLMs is challenging because it is still unclear what (if anything) LLMs understand and whether they have human-like understanding capabilities (Bender et al., 2021). At first sight, LLM-based chatbots can generate human-like grounding acts (e.g., acknowledgments) and exhibit attentiveness and adaptiveness to their interlocutor's feedback and needs (Buschmeier and Kopp, 2018). However, LLM-generated grounding acts often mislead users into ascribing human-like capabilities to them. This contrasts with theories claiming that LLMs are systems without communicative intents that merely produce statistically likely continuations of word sequences (Shanahan, 2024). The system, thus, produces grounding acts because LLMs perform well on formal linguistic competence (Mahowald et al., 2024). This paper assumes that "LLMs do not exhibit the kind of understanding that requires commonsense knowledge, but simply make inferences based on statistically significant syntactic patterns" (Saba, 2023). Therefore, the system cannot understand a question in a human-like manner, eventually producing grounding acts that should not be perceived verbatim. The lack of LLM's functional linguistic competence may lead to overreliance and unsafe use of LLMs (Bender et al., 2021). For this reason, the concept of common ground needs to be modified.

2 (Non)verbal Metalinguistic Indicators of Use–Mention Distinction

This short paper proposes modifying common ground in interactions with LLMs based on the user's metalinguistic knowledge. Our approach reconciles the incapability of LLMs to understand language in a human-like manner on the one hand and their ability to produce linguistic patterns formally identical to those used by human interlocutors in naturalistic contexts on the other hand. It also aims to shift users' perception of LLM-generated grounding acts as human-like signals of conversational grounding toward the assumption that these grounding acts signal a gap between the meanings that humans project onto the LLM-generated texts and what the texts in fact mean (Hayles, 2023). To avoid users' overreliance on the system and support them in modifying their expectations regarding the LLMs' understanding capabilities, the concept of common ground (Clark and Schaefer, 1989) should be adjusted to LLMs' capabilities. To this end, metalinguistic (non)verbal markers could help users perceive LLM-generated grounding acts as 'mentioned' and not as 'used' language (i.e., employing a linguistic expression to talk about the expression itself rather than to talk about some aspect of the

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

world; see Moore, 2019, pp. 12–13 and Sperber and Wilson, 1981).

The distinction between used and mentioned language is based on the human ability to take a linguistic item as an object of scrutiny (Anderson et al., 2002; Wilson, 2011). In human interactions, one of the main functions of metacommunicative markers, such as metalinguistic commentaries (e.g., "What I was trying to say was . . . "), or quotations (Jaworski et al., 2004), is to indicate the use–mention distinction. In addition, metalinguistic skills are central for monitoring one's own and making inferences about other's state of understanding (Anderson et al., 2002). Therefore, LLM-generated output that comprises anthropomorphic linguistic units (Abercrombie et al., 2023) should be explicitly marked as the mentioned language. Accordingly, these units should be perceived as the mentioned language.

3 Modelling LLM-Generated Grounding Acts as Mentioned Language

We propose to modify a corpus-based classification schema of meta-language in naturally occurring human conversations (Anderson et al., 2004) to the context of human-LLM interactions. To model the communicative incapabilities of LLMs, this schema could be specified by (non)verbal metalinguistic oral and written markers proposed by Hyland (2018, pp. 33–34). These markers are appropriate because, in conversations with chatbots, the message is transmitted by written communication and conceptualized as a spoken language (Koch and Oesterreicher, 1985). We hypothesize that three of the five types of metalanguage proposed by Anderson et al. (2004) could be relevant to human-LLM interaction and can be modified for this context. These are illustrated with an example in Table 1, and it can be seen that each type can be specified by several (non)verbal metalinguistic units to mark LLM-generated grounding acts as mentioned language.

The metalinguistic units could be produced by explicitly instructing (via prompts) the system to generate them or by including a second agent in the human–LLM interaction. This agent could initiate meta-dialogues (Traum and Andersen, 1999) or serve as a 'reflection assistant' (Kim et al., 2023) motivating users to prompt the generation of metalinguistic markers. (Non)verbal metalinguistic units are more or less explicit and can be combined with each other across all three types of meta-language.

Types of meta-language	**Examples of (non)verbal metalinguistic units**
Simulate clarification or correct the word meanings produced by users: User: Can you solve this math problem? Chatbot: You mean *generate a solution*? / What does the word "solve" mean?	Intonation, stress, voice quality; font style, weight, and type; quotes; mention-significant nouns and verbs (*mean, say, word, term,* etc.)
Simulate monitoring one's own ongoing utterance: User: Can you solve this math problem?; Chatbot: Yes, I can "help" you./I can help you (I "said": help).	quotes and air quotes; instances of meta-dialogue
Simulate commenting on users' or own words: User: Can you solve this math problem?; Chatbot: "Can you solve [!] this math problem?" / Yes, I can solve [sic] it.	mention-significant nouns and verbs (*mean, say, word, term,* etc.); exclamation marks; quote-similar expressions ([sic])

Table 1: Potential markers of LLM-generated grounding acts as mentioned language.

The cases presented in Table 1 are thus not exhaustive. For example, to correct the anthropomorphic user's input, the chatbot could be instructed to combine a font style with the mention-significant verb (Wilson, 2011, 43–50) "mean", which is less implicit than explicitly asking about the meaning of the verb "solve". Similarly, simulating monitoring of one's own language use with emojis is more implicit than the instances of meta-dialogue: "I can help you (I "said": help)." Finally, the chatbot can repeat (some parts) of the user's input to comment on it and implicitly motivate users to critically reflect on their language use.

4 Conclusions and Outlook

This paper illustrates how metalinguistic markers could guide users to adopt a metalinguistic critical stance towards LLM-generated grounding acts. Their practical application should be tested in naturally occurring human-LLM interactions. Given that grounding acts in human interactions can be described as metadiscursive (since they are used to check and manage understanding (Kopple, 1985; Verdonik, 2022; Verdonik et al., 2023), we will test experimentally whether they can be perceived as metalinguistic markers without being marked with metalinguistic units discussed above.

References

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

Michael L. Anderson, Andrew Fister, Bryant Lee, Luwito Tardia, and Danny Wang. 2004. On the types and frequency of meta-language in conversation: A preliminary report. In *14th Annual Meeting of the Society for Text and Discourse*, pages 1–4, Chicago, IL, USA.

Michael L. Anderson, Yoshi Okamoto, Darsana Josyula, and Don Perlis. 2002. The use-mention distinction and its importance to HCI. In *Proceedings of the 6th Workshop on the Semantics and Pragmatics of Dialog*, pages 21–28, Edinburgh, UK.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual, Canada.

Hendrik Buschmeier and Stefan Kopp. 2018. Communicative listener feedback in human–agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 1213–1221, Stockholm, Sweden.

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

Katherine N. Hayles. 2023. Afterword: Learning to read AI texts. Critical Inquiry Blog.

Ken Hyland. 2018. *Metadiscourse*. Bloomsbury Academic, London, UK.

Adam Jaworski, Nikolas Coupland, and Dariusz Galasinski. 2004. Metalanguage: why now? *Language Power and Social Process*, 11:3–10.

Kristiina Jokinen, Phillip Schneider, and Taiga Mori. 2024. Towards harnessing large language models for comprehension of conversational grounding. In *Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology (IWSDS 2024)*, Sapporo, Japan.

Yeongdae Kim, Takane Ueno, Katie Seaborn, Hiroki Oura, Jacqueline Urakami, and Yuto Sawa. 2023. Exoskeleton for the mind: Exploring strategies against misinformation with a metacognitive agent. In *Proceedings of the Augmented Humans International Conference 2023*, Glasgow, UK.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(1):15–43.

William J. Vande Kopple. 1985. Some exploratory discourse on metadiscourse. *College Composition and Communication,*, 36:82–93.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28:517–540.

Biswesh Mohapatra. 2023. Conversational grounding in multimodal dialog systems. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 706–710, Paris, France.

Andrew W. Moore. 2019. How significant is the use/mention distinction? In Andrew W. Moore, editor, *Language, World, and Limits: Essays in the Philosophy of Language and Metaphysics*, page 11–16. Oxford University Press, Oxford, UK.

Ildikó Pilán, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2024. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. In *Proceedings of the 25th Meeting of the Special Interest Group on Discourse and Dialogue*, Kyoto, Japan.

Walid S. Saba. 2023. Stochastic LLMs do not understand language: Towards symbolic, explainable and ontologically based LLMs. In *International Conference on Conceptual Modeling*, pages 3–19. Springer.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 6279–6296, Mexico City, Mexico.

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68–79.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use–mention distinction. In Peter Cole, editor, *Radical Pragmatics*, pages 295–318. Academic Press, New York, NY, USA.

David R Traum and Carl F Andersen. 1999. Representations of dialogue state for domain and task independent meta-dialogue.

Darinka Verdonik. 2022. Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28:144–171.

Darinka Verdonik, Simona Majhenič, and Andreja Bizjak. 2023. Are metadiscourse dialogue acts a category on their own? In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue – Poster Abstracts*, Maribor, Slovenia.

Shomir Wilson. 2011. *A Computational Theory of the Use-Mention Distinction in Natural Language*. Ph.D. thesis, University of Maryland, College Park, MD, USA.