Evaluating LLMs Through Self-Play: Testing Comprehension and Generation of Referring Expressions in Multi-Modal LLMs by Employing a Picture-Guessing Dialogue Game

Eileen Kammel, Anne Beyer, David Schlangen (University of Potsdam)
eileen.niedenfuehr@uni-potsdam.de

Referring expressions (REs) are integral to communication, as they allow speakers to identify and refer to specific entities within discourse. Human speakers tend to adapt and optimize their expressions based on context and the intended referent by adhering to conversational norms (Grice, 1975). This study examines how large language models (LLMs), particularly multi-modal models, approach this task, investigating their success and limitations in processing REs in different contexts. The research employs a version of the picture-guessing game *reference-game* as implemented in the clembench framework (Chalamalasetti et al., 2023). Using two image sets of varying levels of complexity, key questions explored in this work include (1) whether LLMs scores differ in RE production versus RE comprehension, (2) whether the complexity of images impacts performance, and (3) how the amount of information in LLM-generated REs compares to both theoretically optimal REs and those produced by human speakers. Challenging the latest SOTA open-source models and selected commercial LLMs, the study allows for evaluation of possible performance discrepancies between models of the open-source versus the commercial tier as well as a comparison between models of the same tier.

The game involves two players presented with the same images in different orders. One player describes a target image to differentiate it from the others, while the second player must select the correct image based on this description. Attributes may be shared between the images, making context important. Images from the TUNA corpus (Gatt et al., 2009) are chosen for simpler images, 3D shapes corpus (Burgess and Kim, 2018) provides more complex images. Following Glucksberg et al. (1966), who have shown that the development of understanding and production of REs can be attested to different phases during language acquisition, we also investigate the presence of these abilities in LLMs separately.

To assess comprehension, successful REs are collected from human trials and presented to an LLM acting as a guesser. Pairs of volunteers play the game via the chat-room like slurk framework (Götze et al., 2022). Expressions that led to successful target identification at least once are presented to the LLMs. To rule out potential location biasthe images are shown in all possible permutations, and the model’s responses are compared across these variations.

For production assessment, the LLM generates the REs. To enable a more nuanced evaluation, context-dependent descriptions are compared to “ground truth” descriptions generated by the same models without context for the same target image. This comparison aims to determine whether contextual information influences the LLM’s descriptions. Additionally, successful game episodes are compared across both comprehension and production tasks to uncover potential differences in the challenges posed to LLMs by these two settings. Building on the findings of Hakimov et al. (2024), who noted that LLM-generated descriptions often resemble image captions with excessive detail, prompts will be adapted to elicit more concise, RE-like expressions, adhering to communicative maxims described by Grice. In the qualitative analysis, expressions generated by both humans and LLMs will be compared to optimal expressions calculated using methods such as the brevity algorithm and the incremental algorithm, as described by Dale and Reiter (1995).

References

Burgess, C. and Kim, H. (2018). 3d shapes dataset. https://github.com/google-deepmind/3d-shapes.

Chalamalasetti, K., Götze, J., Hakimov, S., Madureira, B., Sadler, P., & Schlangen, D. (2023). clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents. In H. Bouamor, J. Pino, & K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 11174–11219). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.689

Dale, R., & Reiter, E. (1995). Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. In Cognitive Science (Vol. 19, Issue 2, pp. 233–263). Wiley. https://doi.org/10.1207/s15516709cog1902_3

Gatt, A., Belz, A., & Kow, É. (2009, March 1). The TUNA-REG Challenge 2009: Overview and Evaluation results. ACL Anthology. https://aclanthology.org/W09-0629

Glucksberg, S., Krauss, R. M., & Weisberg, R. (1966). Referential communication in nursery school children: Method and some preliminary findings. Journal of Experimental Child Psychology, 3(4), 333–342. https://doi.org/10.1016/0022-0965(66)90077-4

Götze, J., Paetzel-Prüsmann, M., Liermann, W., Diekmann, T., & Schlangen, D. (2022). The slurk Interaction Server Framework: Better Data for Better Dialog Models. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4069–4078). European Language Resources Association. https://aclanthology.org/2022.lrec-1.433

Grice, H. P. (1975). Logic and Conversation. In Speech Acts (pp. 41–58). BRILL. https://doi.org/10.1163/9789004368811_003

Hakimov, S., Abdullayeva, Y., Koshti, K., Schmidt, A., Weiser, Y., Beyer, A., & Schlangen, D. (2024). Using Game Play to Investigate Multimodal and Conversational Grounding in Large Multimodal Models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2406.14035