Are Large Language Models capable of Gricean conversational implicatures?

Hélie Bazin*[1]

[1]Sorbonne Université - Faculté des Lettres - UFR de Philosophie – Sorbonne Université - Faculté des Lettres, so – France

Résumé

Recent empirical works on the linguistic abilities of Large Language Models (LLMs) are said to be demonstrating the capabilities of LLMs to detect implicit meaning in conversations,
notably with Gricean conversational implicatures. From those empirical results, many authors
claim that LLMs understand the pragmatic aspects of our language. Drawing on a careful analysis of Gricean conversational implicatures as well as recent works in the epistemology of
understanding and philosophy of AI, I question this claim and offer a reflection on what it would take for LLMs to succeed in those pragmatic tasks.
I first make a quick review of LLMs' architecture, supervised learning, prompting techniques and reinforcement learning with human feedback (RLHF), as well as the overall working of Grice (1957, 1975)'s theory for conversational implicatures. I present Kim, Taylor, and Kang
(2023)'s paper in which the authors confront the ability of different LLMs to detect implicit utterances in conversational scripts to those of a human panel. LLMs get overall better scores which makes the authors claim that LLMs understand conversational implicatures. I also mention studies from Ruis et al. (2023), Qui et al. (2023) and Bojic et al. (2023) with similar results and conclusion.
I then challenge the claim that LLMs understand conversational implicatures using recent studies in the epistemology of understanding, mainly from Pritchard (2009). I argue that LLMs lack the intersubjective component of speech required for conversational implicatures.
Finally, I review current debates in the philosophy of AI (Millière 2024a, 2024b) and argue that training LLMs with a
single reward function is insufficient for systems with goal-directed behavior, which is needed for implicit speech.
PARTIAL BIBLIOGRAPHY
Bojic, L., Kovacevic, P., & Cabarkapa, M. (2023). GPT-4 Surpassing Human Performance in Linguistic Pragmatics. arXiv preprint arXiv:2312.09545.

Grice, H. Paul (1975). Logic and Conversation. In Donald Davidson (ed.), The logic of grammar. Encino, Calif.: Dickenson Pub. Co.. pp. 64-75.

Grice, Herbert Paul (1957). Meaning. Philosophical Review 66 (3):377-388.

*Intervenant

Kim, Z. M., Taylor, D. E., & Kang, D. (2023). " Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures. arXiv preprint arXiv:2305.13826.

Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models– Part I: Continuity With Classic Debates. arXiv preprint arXiv:2401.03910.

Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models-Part II: The Way Forward. arXiv preprint arXiv:2405.03207.

Pritchard, Duncan (2009). Knowledge, Understanding and Epistemic Value. Royal Institute of Philosophy Supplement 64:19-43.

Qiu, Z., Duan, X., & Cai, Z. G. (2023). Pragmatic implicature processing in ChatGPT.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2023). The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. Advances in Neural Information Processing Systems, 36, 20827-20905.

Mots-Clés: IA, Grice, conversational implicatures, understanding