Are conversational large language models speakers?

Paul Piwek
The Open University, United Kingdom
paul.piwek@open.ac.uk

> *Fundamental understanding, you can hardly argue with that.*
> Kees van Deemter (van Deemter and Mineur, 1994, 58)

With the advent of large language models (LLM), and in particular their framing as chatbots – that is, conversational agents – the original and time-honoured test for determining whether machines can think, the Turing test (Turing, 1950), has been called into question. We have reached a point where current generations of conversational LLM can pass time-limited versions of the test (Jones and Bergen, 2023). Additionally, the very ability of machines to pass the test is no longer considered to be a genuine indicator of thinking, though it may be a good indicator of the capability for deception (Biever, 2023).

Recently, informal arguments, such as the Octopus test thought experiment (Bender and Koller, 2020) have been put forward purporting to show that systems that are trained only on (language) form cannot understand language. In this paper we will refrain from taking a stance on this argument, and instead raise a further question which considers conversational LLMs from the point of language generation or production rather than understanding. The question we aim to address is: 'Are large language models speakers?' Conversational LLM have brought back to attention fundamental questions about what it means to be a language user and, in line with the quote at the beginning of this paper, we believe this is a good thing.

We start by considering the foundational contribution to linguistic pragmatics made by H.P. Grice (Grice, 1957). Grice investigated what is involved in a speaker meaning something when they use language. In fact, Grice subsumes speaker meaning under, what he calls, non-natural meaning, in contrast with natural meaning. As examples of natural meaning, Grice provides regularities in nature such as smoke meaning fire and a rash meaning measles. Grice proposes that non-natural meaning is fundamentally different from natural meaning. As an example of a situation involving non-natural meaning, Grice asks us to consider that three rings on a bus, at the least in England at the time Grice wrote his paper, meant non-naturally (meant$_{NN}$) that the bus is full. As a first approximation, Grice suggests that such an 'utterance' u has a non-natural meaning if it was intended by its utterer to induce a belief in some 'audience'. Grice then proceeds to refine this description of non-natural meaning by considering cases that reveal the shortcomings of this first approximation: 'I might leave *B*'s handkerchief near the scene of a murder in order to induce the detective to believe that *B* was the murderer; but we should not want to say that the handkerchief (or my leaving it there) meant$_{NN}$ anything or that I had meant$_{NN}$ by leaving it that *B* was the murderer.' (Grice, 1957, 381-382) After further rounds in which Grice considers other limitations of the initial formulation, he eventually arrives at the proposal that *A* meant non-naturally something is equivalent to *A* uttered u with the intention of inducing a belief by means of the recognition of this intention.

Gricean non-natural meaning allows us to characterise speakers as producers of non-natural meanings. The definition does however assume a prior understanding of the notions of belief, intention and recognition. It is tempting to interpret these as psychological states or processes. However, the treatment of such folk psychological notions as foundations for science has been criticised from various angles, e.g., by problematising the concept of belief as foundation for cognitive science (Stich, 1983) and our common sense understanding of conscious experiences (Frankish, 2016). Similarly, the notion of intentions or psychological reasons has not escaped scrutiny: 'Why do you think this? Why did you do that? We answer such questions by giv-

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

ing reasons, as if it went without saying that reasons guide our thoughts and actions and hence explain them. (...) It is based, however, on a convenient fiction: most reasons are after-the-fact rationalizations.' (Mercier and Sperber, 2017, 109)

Returning to the topic of conversational LLMs, it is also not clear how to apply these folk psychological concepts to conversational LLMs. It seems somewhat too convenient to simply dismiss the possibility of conversational LLMs as speakers on the basis that they don't have intentions or goals. It is not prima facie clear that they completely lack intentions or at least functionally equivalent states. Though LLM training (i.e. pretraining) is limited to the next word prediction task, conversational LLMs are finetuned in ways that arguably do instill implicit goals on how to follow instructions and avoid inappropriate responses (Ouyang et al., 2022). Furthermore, explicit user prompts or hidden system prompts/context could also be argued to introduce goals.

To be fair to Grice, he specifically writes that he does not want to 'peopl[e] all our talking life with armies of complicated psychological occurrences' (Grice, 1957, 386) and gestures at what is 'normally conveyed', 'refer[ence] to the context', and 'asking the utterer afterward' (Grice, 1957, 387). This line of thought is suggestive of an alternative approach to the question whether conversational LLMs are speakers grounded in a view of language use as participation social practices or Wittgensteinian language games (Wittgenstein, 1953).

A potentially fruitful twist to this approach is proposed by Robert Brandom (Brandom, 1994, 2000), who works out in detail how the language game of giving and asking for reasons is fundamental to all other language games in that this specific game explains the representational power of language - i.e. the language – world relationship. Doing so, he espouses an unusual explanatory move from pragmatics to semantics.

In a nutshell, the game of giving and asking for reasons – for partial formalisations see (Kibble, 2006; Piwek, 2011, 2014) – puts certain normative demands on interlocutors, in particular, an assertion (e.g., 'It rains') results, downstream, in commitments (e.g. prohibiting inconsistent assertions such as 'It doesn't rain' or 'It snows') and, upstream, in potential challenges about the entitlement to or justification for that assertion ('The tiles wet.').

Mastery of this game of giving and asking for reasons may provide us with some insight into the extent to which conversational LLMs are speakers. Interestingly, in as far as commitments and consistency are concerned, conversational LLMs have and continue to struggle with negation (e.g. tests with the prompt 'I do not have two apples. I give one away. How many apples do I have?') causes chatGPT to produce correct responses about 3 out of 5 times, but also bizarre incorrect ones such as 'You have on apples left (...)' (ChatGTP4o, 5 July 2024). Testing Gemini and ChatGPT4o for their way of dealing with contradictions – i.e. challenging its assertions – we found that, after challenging the result of calculating the product of two large numbers, Gemini always concedes that the user is right (even if they clearly aren't) whereas ChatGPT4o, after each challenge, responds with 'To ensure absolute accuracy, I will recompute once again'. Both are appropriate machine responses, but nothing like the behaviour of a speaker who cares about their contribution to the conversation and is sensitive the assessment by others.

This final point is fundamental, resting on the view of speaking (**S**) as a contribution by a person to a language game, i.e. a normative social activity requiring (i) sensitivity to, i.e. caring about, peer assessment of one's contributions and (ii) engagement with peer assessment of others' contributions.

In contrast, automatic natural language generation (**A**) is the algorithmic generation of output strings that we take to be English or French or Chinese or ..., given a (more or less formal) specification of requirements on the output (e.g. a prompt, logic formula or other).

We'd like to conclude by proposing that the current perspective on speaking and generation raises both a concern and challenge. Let's start with the concern, which can be seen as our variation, and attempt at clarification, of the Eliza effect (Weizenbaum, 1966) and the more general Media Equation (Reeves and Nass, 1996):

The chatbot conceit = *the design of systems that do* **A** *but appear to be in the business of doing* **S** *by framing interactions as dialogue.*

On the positive side, for researchers in pragmatics a daunting but also invigorating challenge remains and has, arguably, been rekindled by the recent advent of conversational LLM:

The pragmatics challenge: *What are the ingredients* **I** *such that* **A** + **I** = **S***?*

Acknowledgments

The argument presented in this paper was originally prepared for an informal gathering on the 6^{th} of July 2024 in honour of Kees van Deemter, who I'd like to thank for the many discussions we've had and will hopefully have in future about meaning, language generation and many other things. The way Kees succeeds in caring for and combining both open-mindedness and rigour exemplifies to me what it means to be a speaker in the sense of this short paper.

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Celeste Biever. 2023. ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature*, 619:686–689.

Robert Brandom. 1994. *Making It Explicit: reasoning, representing, and discursive commitment*. Harvard University Press, Cambridge, Massachusetts.

Robert Brandom. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Harvard University Press, Cambridge, Massachusetts.

Keith Frankish. 2016. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11-12):11–39.

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

Cameron Jones and Benjamin Bergen. 2023. Does GPT-4 Pass the Turing Test? *arXiv preprint*. ArXiv:2310.20216 [cs].

Rodger Kibble. 2006. Reasoning about propositional commitments in dialogue. *Research on Language and Computation*, 4(2-3):179–202.

Hugo Mercier and Dan Sperber. 2017. *The Enigma of Reason*. Harvard University Press.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*. ArXiv:2203.02155 [cs].

Paul Piwek. 2011. Dialogue structure and logical expressivism. *Synthese*, 183(1):33–58.

Paul Piwek. 2014. Towards a computational account of inferentialist meaning. In *Proceedings of the 50th Anniversary Convention of the AISB*.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press.

Stephen Stich. 1983. *From folk psychology to cognitive science: The case against belief.* The MIT Press.

A. M. Turing. 1950. Computing machinery and intelligence. *Mind*, 59(236):433–460.

Kees van Deemter and Anne-Marie Mineur. 1994. Kees van Deemter interviewed by Anne-Marie Mineur: "Fundamental begrip, daar kun je bijna niet tegen zijn.". *Ta! studentenblad computationale taalkunde*, 4(2):58–69.

Joseph Weizenbaum. 1966. Eliza a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Ludwig Wittgenstein. 1953. *Philosophical investigations. Philosophische Untersuchungen*. NY, Macmillan. (1953). x, 232 pp.