Large Language Models and the Philosophy of Language Games

Abstract

Recent discussions about the performances of large language models (LLMs) have revived philosophical interest in possible interpretations of syntactical and semantical aspects. There is a growing stream in the literature relying on Wittgensteinian language games, which can serve as a conceptual framework for non-referential approaches. This paper will go back to the works of Wilfrid Sellars and Kristóf Nyíri in order to analyze the conditions under which LLM based AI-systems might be said to "understand" language. Finally, emphasis is placed on the role of dialogical interactions in social grounding as the relevant background for pragmatic semantics.

1. Large language models and linguistic understanding

In a recent contribution, Tom Schaul (2024) of Google DeepMind presented a framework for understanding large language models (LLMs) based on the Wittgensteinian concept of *language games*. This approach was also reflected in one of the earliest reports on experiments with LLM-based chatbots by Blaise Agüera y Arcas (2022) from Google Research, where he documented and analyzed his initial interactions with the LaMDA model.

Simultaneously, contemporary discussions surrounding LLM performance have revived enduring epistemological and metaphysical questions, such as whether an AI system can be said to truly "understand" anything. In the context of grounding semantics, Holger Lyre (2024) sees three possible dimensions—functional, social, and causal—, and he argues that LLMs exhibit foundational evidence in all three.

In his systematic overview Lyre outlines five methodological options, distinguishing between external (third-person) and internal (first-person) approaches. The standard extrospective-behavioral method is represented by the Turing Test (E1), which he contrasts with top-down approaches rooted in semantic theory (E2) and bottom-up analyses focused on mechanisms that generate semantic capabilities (E3). He finds little merit in introspective methods, such as self-reports (I1) or self-explanations (I2). While E1 and I1 are empirically underdetermined, thus leading to concerns about the so-called "semantic zombie"; E3 is hindered by the "black box" problem, which has led some to see mere "stochastic parrots".

Following the way of E2, Lyre identifies three key theoretical traditions in philosophy of mind and language with implications for semantic grounding: causal-teleosemantic theories, use theories, and functional role semantics. The first two are externalist, while the third is internalist. Each provides a distinct criterion contributing to the feasibility of semantic grounding. This results in a threefold distinction: (1) causal grounding, (2) social grounding, and (3) functional grounding. Whereas the third is entirely internal—lacking interaction with the external world—the first two involve real-world engagement through linguistic behavior.

This will already recall Wittgenstein saying that meaning is nothing else than the functional roles played by the given utterance in a language game. Also, this use of language is necessarily presented in a form of a social practice.

In a very much similar vein, Reto Gubelmann (2022) argues that in some sense current transformer-based NNLP models are close to "understand" language. As he formulates it: „Word embeddings […] are systematic, mathematically powerful ways to register the use of words. The relationships represented in these embedding spaces are emphatically not restricted to the syntactical domain […]." (Gubelmann, 2022: 17)

While other approaches seem to expect too much of a being for attributing language understanding to it, from a „Wittgensteinian-deflationary" perspective we may think that „whether a being understands language depends on the being's competencies, as they are evident in its autonomous adaptability to and

performance of a wide variety of linguistic tasks in a multitude of different settings, as it is exemplified in the task-switching inherent in the Turing Test." (Gubelmann, 2022: 2)

2. Language games for non-referential semantics

Nevertheless, this way of thinking is anything but new. As J. C. Nyíri (1989: 389) says in a footnote in his paper on the philosophical understanding of AI: "In my paper 'No Place for Semantics' […] I have tried to show how the approach of Sellars, and of course that of Wittgenstein, makes the distinction between syntactic and semantic rules lose all its naturalness. I take this to be one of the many points where John Searle's ill-reputed Chinese room argument breaks down […]".

In the above mentioned paper Nyíri (1971) critiques traditional theories that ground meaning in denotation—whether to external reality, mental states, or both—arguing instead for an immanent description of language in use. He finds such a perspective in Wittgenstein and Sellars. His wording here resonates strongly with descriptions of contemporary LLMs: „The meaning of a word is constituted by its role – or, to formulate it in a somewhat different, for our present purpose more suitable way: the meaning of a word is determined by the place which it occupies in the totality of language, i.e., in the totality of possible sentences." (Nyíri, 1971: 57)

The investigations of Ludwig Wittgenstein (1953/1958) in the nature of meaning can be especially useful if one is searching for explanations of linguistic phenomena without relying on the notions of denotation and reference. This may also help to avoid drowning in mentalistic terms when talking about linguistic behaviors of various AI systems.

Wilfrid Sellars (1954) himself had his own understanding of the phenomena called language games, which can also be informative for these discussions. He does not aim to provide a detailed psychological account of how organisms learn pattern-governed behavior but rather to make plausible the idea that an organism can participate in a language game—moving from position to position, doing this just "because of the system," without invoking any kind of metalanguage games.

Sellars (1954: 210) suggests that we shall make some distinctions among the transitions in a language game. We shall take apart (1) *moves*, where both *stimulus* and *response* are internal to the system, from (2) another type, which somehow transcend the world of the language game. The latter is further specified into (2.1) *language entry transitions*, where the stimulus would be said "to be meant by" the response, and (2.2) *language departure transitions*, where the response would be said "to be meant by" the stimulus. The detailed examination of these cases in connection with LLMs can lead to useful insights on their epistemological and metaphysical statuses.

3. Artificial intelligence and dialogical construction of meaning

In this framework, semantic preciseness arguably increases with enrichment of context. In a lack of that, utterances produced by LLMs will show mediocrity, as Agüera y Arcas (2022: 185) already alerted: „LaMDA is indeed, to use a blunt (if, admittedly, humanizing) term, bullshitting. That is because, in instructing the model to be sensible and specific–but not specific in any specific *way*–bullshit is precisely what we have requested. The model has no instinctive or acquired preferences the way we do; nor does it have a body, or senses, or any narrative recollection of an autobiographical past." What follows from that is that the challenge of increasing semantic depth will be of fundamental importance in enhancing LLMs performances, whether it will happen via targeted pre-training, sophisticated fine-tuning, or real-time access during their operation to external resources.

From the perspective of the philosophy of language games, however, the real issue should be social usage. How can we locate this aspect in the making of meaning in the cases of LLMs? Instead of trying to locate it in some of their inner layers, there is a trivial way to find those groundings in accordance with the above theoretical framework. To do that we shall only make a move from bare LLMs to dialogical AI-services built around them–conversations going on between human users and AI systems will provide the necessary semantics.

In these interactive settings the construction of meaning becomes a shared responsibility, mirroring the inherently collaborative dynamics of human dialogue. Errors—whether originating from human

participants or machine agents—are to be expected; still, what is of real significance is the dialogical process that shall contain mechanisms capable of resolving such misunderstandings, thereby enabling progression toward mutual understanding. All players of the language game will typically share the overarching aim of achieving a joint comprehension of the issues currently under discussion. As Agüera y Arcas (2022: 185–186) observed: „This offers us a clue as to why mutual modeling is so central to dialogue, and indeed to any kind of real relationship. Hiding behind the seemingly simple requirement for interlocutor *A* to remain consistent in its interactions with *B* is an implication that *B* is modeling *A* (so, will notice an inconsistency), thus the requirement for *A* not only to model *B*, but to model *B*'s model of *A*, and so on." It underscores the inherently intersubjective nature of conversational engagement, wherein each participant actively anticipates and adjusts to the evolving mental models of the other, fostering a dynamic and responsive communicative process.

In conclusion it appears that the work of semantics will get definitely done only by engaging LLMs in dialogue with human interlocutors utilizing so-called chatting technologies. From the perspective of the language game paradigm, these added functions will not be mere auxiliary or entertaining features. Rather, they constitute essential components of these communicative systems, enabling them to move beyond abstract symbol manipulation and to reach out to the real world.

References

Agüera y Arcas, B. (2022). Do Large Language Models Understand Us? *Daedalus* 151, 183–197.

Gubelmann, R. (2022). A Loosely Wittgensteinian Conception of the Linguistic Understanding of Artificial Neural Networks. *Grazer Philosophische Studien* 99, 4, 485–523.

Lyre, H. (2024). "Understanding AI": Semantic Grounding in Large Language Models. *arXiv*:2402.10992.

Nyíri, J. C. (1971). No Place for Semantics. *Foundations of Language* 7, 1, 56–69.

Nyíri, J. C. (1989). Wittgenstein and the Problem of Machine Consciousness. *Grazer Philosophische Studien* 33, 1, 375–394.

Schaul, T. (2024). Boundless Socratic Learning with Language Games. *arXiv*:2411.16905v1.

Sellars, W. (1954). Some reflections on language games. *Philosophy of Science* 21, 3, 204–228.

Wittgenstein, L. (1953/1958). *Philosophical Investigations.* Transl. G. E. M. Anscombe. Oxford: Blackwell.