Application: Master in Computer Science (TUM)

Topic 4: Can computers think?

Sometimes, LLMs Think: Mechanistic Interpretability and Reliable Faculties

Jonas Strabel · May 31, 2026

"either frontier LLMs are capable of sense-making [...], or the kind of linguistic competence they exhibit does not necessarily require sense-making" [1]. While some authors have argued that machines are merely stochastic parrots [2] or symbol manipulators [3], Froese's 'AI dilemma' makes it clear that this is becoming increasingly difficult to maintain. It is hard to imagine that the outputs of modern LLMs do not require any sense-making.

1. Refining the Question

This development in artificial intelligence renders the question, 'Can computers (maybe now) think?', extremely relevant. It is interesting, then, that Alan Turing already proposed to replace the similar question "Can machines think?" with his imitation game [4, p. 433] over 70 years ago, because he thought the question was "too meaningless to deserve discussion" [4, p. 442]. One can see why this might be by looking at what would happen if we were to simply answer it with 'yes' or 'no'. What does that mean? What follows based on that? It is not immediately clear, because the word 'thinking' is ambiguous and the question is highly complex. I do not consider the question to be worthless. Defining what falls under a word and what does not is useful, because that helps us distinguish real differences in the world. But when a word that developed in connection to humans is applied to computers, its criteria are no longer clear. A Wittgensteinian way of putting this would be that the word is being moved into a new language game. To ensure this does not devolve into a fight over words, a meaningful answer is probably not simply a 'yes' or 'no' answer, but a graded and nuanced one [5, p. 3]. It might therefore be more useful to specify sub-questions that use words that are less ambiguous and better defined to use in connection with computers. It also makes sense to choose a question for which it is somewhat clear what explanatory work the attribution is supposed to do.

In this paper, I want to consider the question of whether LLMs have 'reliable faculties' in the sense that the processes with which they generate outputs are based on sophisticated mechanisms that resemble human ones, such as internal representations of concepts that causally affect the output (and not simply cheap shortcuts such as memorization). If that is the case, computers are more useful to humanity because they are probably more reliable, and, I postulate, attributing a certain cognitive aspect of 'thinking' is warranted. This question satisfies the aforementioned criteria for a more useful inquiry and will now be analyzed.

2. Looking Under the Hood

Beckmann and Queloz analyzed the processes of machines, specifically modern LLMs, by looking at empirical studies in Mechanistic Interpretability (MI) that investigate how LLMs generate their outputs. They present "mechanistic evidence that LLMs are capable of forms of understanding

comparable to our own." [6, p. 35]. For example, they define state-of-the-world understanding as involving LLMs "forming an internal model of how distinct features relate to each other to constitute facts" [6, p. 18]. This means that when LLMs output factual sentences, the words are not merely statistically likely but the result of internal connections that mirror the external world. While at the micro-level LLMs operate merely on statistics, at the macro-level, such representational connections seem to emerge. The authors consider an idealized example in which LLMs first combine the features "Golden Gate" and "Bridge" into a single "Golden Gate Bridge" feature. Another learned weight can now connect it to a second combined feature, namely "in San Francisco". When an LLM is queried and the "Golden Gate Bridge" feature activates, this weight causes the "in San Francisco" feature to also activate [6, p. 19]. There are many examples like this in the paper, and they coincide with broader findings such as MI studies by Anthropic [7, 8]. This clearly shows that on the macro-level, LLMs are not merely statistical parrots.

However, LLMs have many different structures and mechanisms by which they come to a result [6, p. 36]. Some of these are sophisticated and resemble human cognition, while others are 'cheaper' statistical or memorized patterns. Although humans also utilize different processes and sometimes take shortcuts, there is still at least one essential difference from what humans do: humans possess meta-cognition that usually allows them to select the right processes for the job because they seem to strive for parsimony [6, p. 37-38]. In contrast, LLMs currently do not seem to reliably do that. One example finding the authors give is the following: Studies found that LLMs have content-based circuits and also formal circuits that correctly process logical structure. However, when presented with a syllogism that is logically valid but nonsensical in content, they are less likely to accept it than one that makes factual sense. The correct and existing circuit gets buried by shallower heuristics [6, p. 38].

3. Conclusion

This paper progressed from the broader question "Can computers think?" to a more answerable sub-question regarding LLMs' "faculties". It has shown that these faculties are indeed sophisticated. Considering their internal workings, it seems that LLMs utilize some processes that one might consider "thinking" in a cognitive sense. While they possess structures and processes that could be considered "thinking patterns", they do not always use the right ones for the job (nor do they reliably realize when they simply do not know), because they seem to lack a sufficiently parsimony-oriented meta-cognition component to do so [6, p. 38]. While still inconsistent, LLMs utilize cognition-like circuits to generate their answers in increasingly many cases. Another way to summarize the current state might simply be: "LLMs can think, sometimes."

However, even if LLMs achieve perfect metacognitive reliability, future analysis must still address the question of intentionality. Several authors argue that real "thinking" or "understanding" requires intention [9, p. 5187][3, p. 417], which computers currently do not seem to possess [10, p. 27-28]. This means that it is too early to conclude that their faculties are consistently reliable or that the word "thinking" can generally be attributed to computers, but it seems to apply to a limited extent when considering the cognitive aspects of the word.

References

[1] Tom Froese. "Sense-Making Reconsidered: Large Language Models and the Blind Spot of Embodied Cognition". In: *Phenomenology and the Cognitive Sciences* (), pp. 1–16. DOI: 10.1007/s11097-025-10132-0.

[2] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? " In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. (Visited on 05/06/2026).

[3] John R Searle. "Minds, Brains, and Programs". In: (1980).

[4] Alan Turing. "Computing Machinery and Intelligence". In: *Mind* 59.236 (1950), pp. 433–60. DOI: 10.1093/mind/lix.236.433.

[5] Holger Lyre. *"Understanding AI": Semantic Grounding in Large Language Models*. https://moodle.sankt-georgen.de/pluginfile.php/47798/mod_resource/content/1/Lyre.pdf. 2024. (Visited on 04/18/2026).

[6] Pierre Beckmann and Matthieu Queloz. *Mechanistic Indicators of Understanding in Large Language Models*. Feb. 2026. DOI: 10.48550/arXiv.2507.08017. arXiv: 2507.08017 [cs]. (Visited on 05/09/2026).

[7] Anthropic. *Tracing the Thoughts of a Large Language Model*. https://www.anthropic.com/research/tracing-thoughts-language-model. (Visited on 04/17/2026).

[8] Anthropic. *Emotion Concepts and Their Function in a Large Language Model*. https://transformer-circuits.pub/2026/emotions/index.html. (Visited on 04/17/2026).

[9] Emily M. Bender and Alexander Koller. "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*. Online: Association for Computational Linguistics, 2020, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. (Visited on 04/18/2026).

[10] Emma Borg. "LLMs, Turing Tests and Chinese Rooms: The Prospects for Meaning in Large Language Models". In: *Inquiry* (Jan. 2025), pp. 1–31. ISSN: 0020-174X, 1502-3923. DOI: 10.1080/0020174X.2024.2446241. (Visited on 04/18/2026).

Declaration of Authorship

I hereby declare that I have written the present essay titled *Sometimes, LLMs Think: Mechanistic Interpretability and Reliable Faculties* independently and have used no sources or aids other than those indicated. All parts of the work taken verbatim or in substance from published or unpublished writings have been identified as such.

Saulheim, May 31, 2026 _____

Location, Date Jonas Strabel