fMRI validation of GPT-4's ability to recognise Theory of Mind in natural conversations

Camilla Di Pasquasio, Rasya Kayleen Tsabitaah, Manon Fleury-Benoit, Marc Cavazza, Thierry Chaminade

[fMRI validation of GPT-4's ability to recognise Theory of Mind in natural conversations]

[Camilla Di Pasquasio[1], Rasya Kayleen Tsabitaah[2], Manon Fleury-Benoit[3], Marc Cavazza[4]*, Thierry Chaminade[1]*]

*equal contribution

[1 Institut de Neurosciences de La Timone, UMR 7289, Aix-Marseille Université—CNRS, Marseille, France]

[2 Faculty of Psychology, Aix-Marseille Université, Aix en Provence, France]

[3 Faculty of Human Sciences, Faculty of Science (MaSCo), Aix-Marseille Université, Marseille, France]

[4 Division of Computing Sciences and Mathematics, University of Stirling, Stirling, FK9 4LA, Scotland, UK.]

Synopsis

Several studies in Artificial Intelligence emphasise the potential of Large Language Models (LLMs) in spontaneously developing Theory of Mind (ToM) due to their exposure to human-like language expressing mental states and reading literary fiction. Recent evaluations of GPT-4's ToM capabilities provide evidence that GPT-4 demonstrates advanced reasoning about the mental states of characters, adept handling of abstract situations, and proposing cooperative actions in social contexts. Here, we introduce the validation of ToM tagging, a tool leveraging LLMs to identify ToM exchanges within transcripts of naturalistic social interactions, using fMRI human brain recordings.

Background

One emerging challenge in Neuroergonomics is the articulation of human cognition with Artificial Intelligence systems. A recent trend in Generative AI, and in particular LLM, has been to investigate their behaviour from a psychological perspective, using tests and methods of cognitive psychology, investigating reasoning [1] or personality traits [2]. There is growing interest in complementing traditional LLM benchmarks with these new behavioural ones [3]. One of these aspects is the ability of LLM to exhibit ToM properties. Although some empirical work has suggested that LLM-based ChatBots could be perceived as empathetic [4], it remains challenging to dissociate such effects from communication aspects in a clinical context. On the other hand, it has been suggested that LLM could be able to exhibit ToM abilities [5], although this has been challenged [6]. Following this, evaluations of GPT-4's ToM capabilities, conducted through tests on both basic and realistic social scenarios, provide substantial evidence supporting the assertion that GPT-4 demonstrates advanced reasoning about the mental states of characters, adept handling of abstract situations, and proposing cooperative actions in social contexts [7].

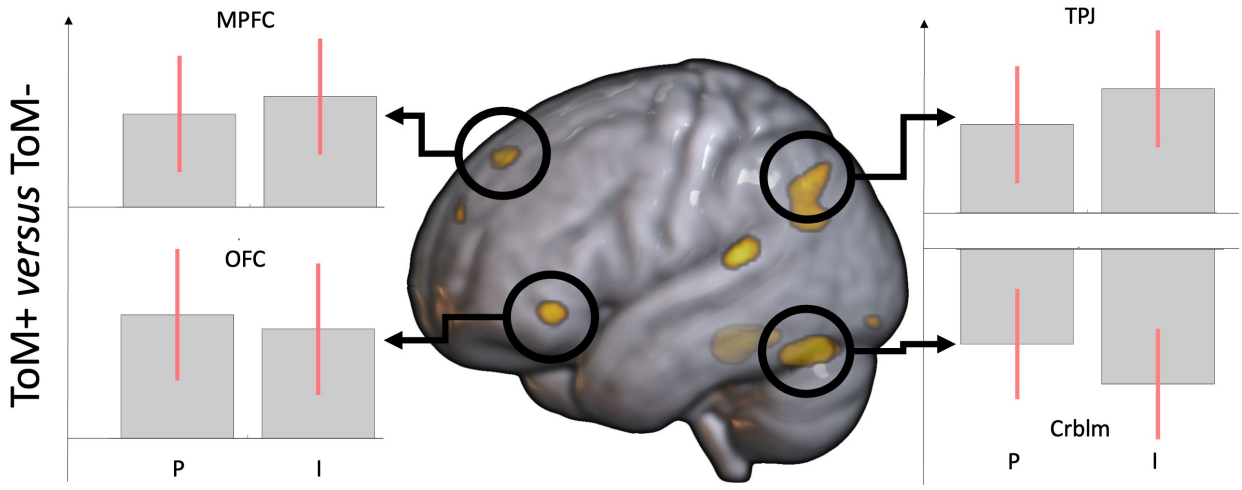
It is important to consider that, despite ToM experiments having taken place in a broader framework of Artificial General Intelligence (AGI) [7] and sometimes referred to as "emergent properties", there exist weaker interpretations such as the acquisition of linguistic expressions of ToM through training corpora that include substantial contribution for novels. The underlying rationale is that novels feature prominently ToM situations, something that has been advocated as reading to develop perspectives in humans [8]. Since there is some evidence that the training corpus of the gpt family incorporates novels since the onset [9], this hypothesis remains valid even at a time when emergent abilities of LLM have been questioned [10]. Since these findings, there has been a growing interest in experimenting how ToM abilities could be related to core aspects of LLM such as Prompting [11] [12]. Furthermore, insight from social cognitive neuroscience studies have extensively explored the cognitive pathways related to ToM, encompassing crucial regions like temporoparietal junction (TPJ), temporal poles (TP), medial

prefrontal cortex (mPFC), and precuneus/posterior cingulate (PCC) , which supports a neuroscience empirical perspective on this controversy [13]. Proposing an innovative approach, we introduce the validation trough neural activation congruent with neuroscientifically validated regions (fMRI) of *ToM tagging*, a tool leveraging LLMs to identify ToM exchanges within transcripts of authentic conversations based on a prior corpus [14]. We suggest employing a dual methodology that leverages Language Models (LLMs) to identify Theory of Mind (ToM) exchanges within transcripts of authentic conversations between two agents.

Methods

We analyse a conversation corpus known to exhibit ToM phenomena [11]. 25 participants recorded with functional MRI while carrying online conversations in 4 sessions x 6 trials alternating between a Human and a Robot interlocutors ("Speakers"). The conversation is considered natural as participants are provided with a cover story hiding the actual objective of the experiment (namely, to study social interactions) and is known to elicit ToM phenomena [12]. We analysed with an LLM the resulting 100 hours of transcribed conversations using a bespoke gpt-4-turbo prompt: each conversation turn from the participant and its interlocutor were tagged as containing (ToM+) or not (ToM-) references to mental states ("Mentalizing") . "Speakers" and "Mentalizing" factors defined the four experimental conditions. The current analysis focused on the contrast ToM+ *versus* ToM- depending on whether the participant or the interlocutor was producing the conversation - *i.e.* whether ToM occurrences were *produced* or *perceived* by the participant. An F-Test of ToM+ *versus* ToM- across the two speakers was used to identify brain regions influenced by the LLM mentalizing categorization for one/both of the speakers ($p_{uncor} < 0.001$, extend > 1 cm^3).

Results



The F-tests identified regions in the right hemisphere temporal parietal and fontal cortices, as well as in the dorsal cerebellum bilaterally. We further explored responses for the contrast ToM+ *versus* ToM- in these regions (see figure). Results indicate that oly the cerebellum clusters are more activated by ToM- than ToM+ (negative estimates) while other regions were more active in ToM+ than ToM- speech turns. Contrast estimates were higher when the participants were producing speech ("P") in the orbitofrontal l cortex, while it was stronger when they were perceiving interlocutors' ("I").

Discussion

The current analysis offer an unique glimpse on the ability of LLM to recognize mentalizing events in natural conversations using fMRI. Results indicate shared neural substrates activated when producing and perceiving mentalizing events in the left medial prefrontal (MPFC),and lateral orbitofrontal cortices (lOFC) as well as temporoparietal junction (TPJ). MPFC and TPJ are classical ToM-associated brain areas, identified through simple observational fMRI experiments, indicating that LLM ToM tagging correctly reflected categorization of behavioural events associated or not with mental state attribution. Furthermore, results suggest that while the MPFC and TPJ respond more to the *perception* of mentalizing events, the lOFC is more strongly involved when participants refer to mental states in their own speech production. Functionally, this OFC is the most anterior part of so-called Broca's area involved in speech production, in which an anterioposterior gradient in which the most anterior area is associated with most abstract representations. Altogether, studying brain correlates of mentalizing in natural conversations offer new insights into how the social brain uses shared neurophysiological correlates to make sense of oneself and other selves verbal production sharing one element, reference to mental states.

References

[1] Dasgupta, I., Lampinen, A.K., Chan, S.C., Creswell, A., Kumaran, D., McClelland, J.L. and Hill, F., 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

[2] Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A. and Matarić, M., 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

[3] Binz, M. and Schulz, E. Using cognitive psychology to understand gpt-3. Proceedings of the National Academy of Sciences, 120(6):e2218523120, 2023.

[4] Ayers, J.W., Poliak, A., Dredze, M., Leas, E.C., Zhu, Z., Kelley, J.B., Faix, D.J., Goodman, A.M., Longhurst, C.A., Hogarth, M. and Smith, D.M., 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, *183*(6), pp.589-596.

[5] M. Kosinski, 'Theory of Mind Might Have Spontaneously Emerged in Large Language Models', 2023, doi: 10.48550/ARXIV.2302.02083.

[6] Ullman, T., 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

[7] S. Bubeck *et al.*, 'Sparks of artificial general intelligence: Early experiments with gpt-4', *ArXiv Prepr. ArXiv230312712*, 2023.

[8] D. C. Kidd and E. Castano, 'Reading Literary Fiction Improves Theory of Mind', *Science*, vol. 342, no. 6156, pp. 377–380, Oct. 2013, doi: 10.1126/science.1239918.

[9] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training.

[10] Schaeffer, R., Miranda, B. and Koyejo, S., 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, *36*.

[11] Tan, F.A., Yeo, G.C., Wu, F., Xu, W., Jain, V., Chadha, A., Jaidka, K., Liu, Y. and Ng, S.K., 2024. PHAnToM: Personality Has An Effect on Theory-of-Mind Reasoning in Large Language

[12] Models. *arXiv preprint arXiv:2403.02246*.

Moghaddam, S.R. and Honey, C.J., 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

[13] Denny, B.T. et al. 2012. A Meta-analysis of Functional Neuroimaging Studies of Self- and Other Judgments Reveals a Spatial Gradient for Mentalizing in Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*. 24, 8 (2012), 1742–1752. DOI:https://doi.org/10.1162/jocn_a_00233.
[14] Chaminade, T., 'An experimental approach to study the physiology of natural social interactions', *Interact. Stud.*, vol. 18, no. 2, pp. 254–275, 2017.