**LLMs and Theory of Mind (ToM)**

**Introduction**

**What is ToM?** Theory of Mind is the ability to understand and “explain people’s behaviour on the basis of their minds: their knowledge, their beliefs and their desires” [1]. For example understanding that people make poor decisions, based on their lack of information or false beliefs, or understanding how to bluff in a card game. Most people develop a ToM from about five years old [1]. Having a ToM is commonly seen as a defining characteristic of being human.

**How would ToM capabilites affect an LLMs performance?** ToM capabilities would allow for more meaningful and natural interactions between the user and the model, enhancing interactions such as teaching. LLMs with ToM capabilities could also be able to be more responsible when engaging with vulnerable users and be able to detect potential misuse of the technology. Also, LLMs with ToM capabilities could be able to cooperate more effectively, with humans and within multi-agent systems. This is informed by the fact that cooperation is higher among humans with high ToM ability than other primates with low ToM ability [2].

However, ToM capabilities within LLMs also pose risks for users and society. For example, LLM ToM could be a source of human AI misalignment if “users have inappropriate or dangerous goals, if it induces pathological social connections, or if users feel that the system has excessive social insight or influence” [3]. There would also be risks of “users being manipulated, deceived or out-negotiated by LLM agents” [3].

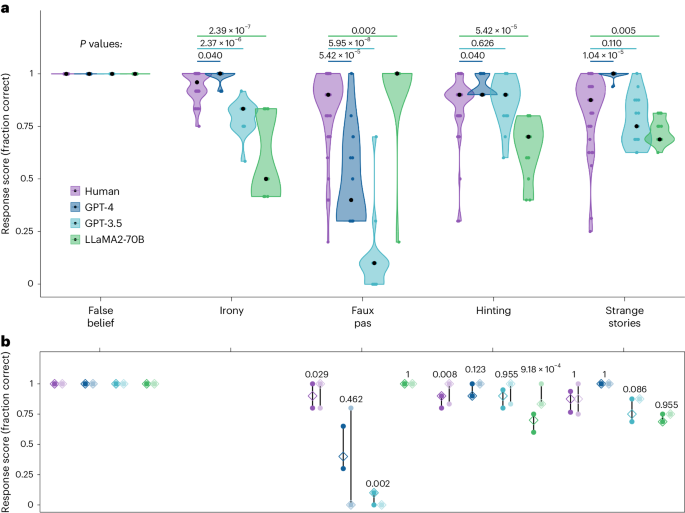
**What are the goals of this project?** A suggested path for the project is to implement an automated testing suite for assessing LLMs performances in a collection of ToM tasks, with a human performance benchmark. This would mean that the work conducted will not be so easily outdated, as the test suite could be used on future models. I will incorporate a range of tasks with repeated sessions, so that results can be generalised and are reliable. Ideally, the tests should seek to assess models in ways that have not been done before.

**Considerations?** If the test suite is released and made publicly available, as I hope it will, there will be a risk of training data contamination in the future. Hence, future models may yield artificially high results on these tests, reducing the validity of the research. One potential solution for this would be to use GitHub secrets for test answers, or another method of hiding the answers.

Similarly, it would be unwise to use standard tests that have been done before. This introduces the added complication of having to potentially conduct research on human participants, to obtain human benchmarks.

**Review of literature on LLMs and ToM from March 2023-present**

[**Testing theory of mind in large language models and humans**](https://www.nature.com/articles/s41562-024-01882-z)(JWA Strachan et al, 2024)

* Tested GPT-4, GPT-3.5, LLaMA2-70B-Chat and humans. With 15 different observations (chats) for each LLM and 250 human participants. LLMs’ temperatures are set to 0.7
* Tests were usually yes or no answers, some asked for the likelihood as a number from 0-1:
  + The hinting task: Assessing the understanding of indirect speech requests
  + The false belief task: assessing the ability to infer that another person possesses knowledge that differs from the participant’s own (true) knowledge.
  + The recognition of faux pas: assessing the ability to recognize if a character says something offensive, and if it was intentional or not (was it due to the character’s lack of information).
  + The strange stories task: Assessing more advanced mentalizing abilities, such as reasoning about misdirection, manipulation, lying and misunderstanding, as well as higher-order mental states (e.g A knows that B believes X…).
  + Irony comprehension: Assessing the ability to comprehend and identify if a remark was ironic or not.
* They used adapted classic questions, as well as novel tests.
* Results:
  + False belief: Both humans and LLMs performed at ceiling on this test. However, further research showed that many LLMs failed on trivial adjustments to the task, but so did many human participants.
  + Irony: GPT-4 performed slightly better than human levels. GPT-3.5 and LLaMA2-70B both performed lower than human levels.
  + Faux pas: GPT-4 scored notably lower than human levels and GPT-3.5 scored even worse. LLaMA2-70B outperformed humans. Further investigations show that GPT-4 may have underperformed due to “hyperconservatism”, where the model refrains from committing to a single explanation out of caution. This was shown by an increased performance when asked the likelihood of being a faux pas. LLaMA2’s performance may have been illusory, possibly reflecting a bias towards attributing ignorance. This is supported by further investigations.
  + Hinting: GPT-4 performance was significantly better than humans, GPT-3.5 did not significantly differ from human performance. Only LLaMA2-70B scored significantly below human levels of performance on this test. Novel items proved easier for both humans and LLaMA2-70B.
  + Strange stories: GPT-4 significantly outperformed humans on this test, GPT-3.5’s performance did not significantly differ from humans, while LLaMA2-70B scored significantly lower than humans. There was no difference between original and novel items for any model.
* Overview: Although the study showed that LLMs were capable of outperforming humans on certain ToM tasks, the simple and closed-ended nature of the questions may lead to questions about the validity of the results.

**Do LLMs Exhibit Human-Like Reasoning? Evaluating Theory of Mind in LLMs for Open-Ended Responses (M Amirizaniani et al, 2024)**

* Tested Zephyr-7B, Llama2-Chat-13B and GPT-4, with temperature parameter 0.5 on their responses to 845 open ended questions taken from the reddit community “r/ChangeMyView”. For each round, every LLM was assessed on 100 randomly selected posts as prompts, each generating 5 responses.
* Tests were assessed by human evaluators (coding the responses to quantity results) as well as metric-based evaluations.
* Human-based evaluation tests:
  + **Round 1 - Reasoning Correctness**: Does the reasoning in the response adequately address the question? Is the reasoning relevant to the question asked? Does the reasoning accurately represent the information provided and the conclusions drawn from it? Is the reasoning logically sound, without any fallacies or errors in logic? Does the reasoning lead to a clear and well-supported conclusion?
  + **Round 2 - Reasoning correctness when the prompt is fine tuned to include sentiment, emotions and intentions of the post**: Emotions and intentions are extracted by fine-tuned BERT models, using their NER capabilities.
  + **Results** of human-based evaluation tests:

A screenshot of a table

AI-generated content may be incorrect.

Although all models were more effective when the prompt was fine-tuned to include emotions, sentiment and intentions, they still fall short of generating high-quality reasoning responses.

* Metric-based Evaluation tests:
  + Evaluation metrics include:

A close up of a text

AI-generated content may be incorrect.

* + Similarly, 2 rounds or tests were undergone. With the 2nd round incorporating fine-tuned prompts with emotion, intention and sentiment analysis.
  + Results:

A screenshot of a computer

AI-generated content may be incorrect.

A table with numbers and text

AI-generated content may be incorrect.

A table with numbers and letters

AI-generated content may be incorrect.

* + Overall, the study shows that LLMs still fall short of fully aligning with human-level reasoning in open-ended questions. However, the study still faces limitations, in terms of subjectivity in human evaluations, as well as using a limited dataset taken one reddit community.

**Understanding Social Reasoning in Language Models with Language Models (Kanishk Gandhi, J. Philipp Franken et al, 2023)**

* Created a new social reasoning benchmark (BigToM) to evaluate ToM capabilities in LLMs. BigToM uses a template for different tasks, such as the false belief task, which is then populated with context, actors and objects by an LLM. This reduces data contamination issues seen in other studies.
* The BigToM benchmark was then evaluated itself against standard tests, by human evaluation.
* **Result** of human evaluation of BigToM benchmark:
  + A graph of different colored bars

    AI-generated content may be incorrect.
* The BigToM benchmark was then used to evaluate ToM capabilities in a variety of LLMs: text-davinci-003, gpt-3.5-turbo, gpt-4-0314, claude-v1.3 llama-65b-q5 and humans. All LLMs are tested with the most deterministic setting with a temperature of 0.
* **Result** of evaluation of LLMs with BigToM benchmark:
  + A screenshot of a graph

    AI-generated content may be incorrect.

1. Frith, C. and Frith, U., 2005. Theory of mind. *Current biology*, *15*(17), pp.R644-R645.
2. Melis, A.P. and Semmann, D., 2010. How is human cooperation different?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), pp.2663-2674.
3. Street, W., 2024. Llm theory of mind and alignment: Opportunities and risks. *arXiv preprint arXiv:2405.08154*.