

Abstract

Recent studies suggest that Large Language Models (LLMs) can perform tasks related to theory of mind (ToM) in ways that are consistent with human behavior. However, these findings have been criticized for their lack of robustness and the foundational claim that human-like cognition and reasoning could arise from language models. Research has found that LLMs are more challenged by the more complex faux pas test than the false belief test. This paper conducts a brief series of A/B tests to evaluate the effectiveness of a Violation of Expectation (VoE) mechanism, implemented through Chain-of-Thought reasoning, in reducing errors in LLM responses to the faux pas test, a clinical psychology method used to test social reasoning. Additionally, the paper critically examines the anthropomorphism of LLMs, suggesting that emergent behaviors might stem from in-context learning or data manipulation, and discusses the ethical implications of machine intelligence that can interact more effectively with humans and anticipate their needs.

Introduction

Large Language Models

Pre-trained language models (PLMs) have demonstrated a remarkable ability to generate grammatically accurate text across a vast array of domains. Large Language Models (LLMs), which are PLMs trained on web-scale data, are able to produce text at a level often indistinguishable from humans. Due to their remarkable capabilities, many findings suggest that these technologies not only represent a significant breakthrough in language processing but also exhibit “sparks of artificial general intelligence” (Bubeck 2023). However, it is important to avoid logical fallacies when evaluating LLMs’ capabilities. Specifically, the tendency to equate linguistic proficiency with cognitive abilities, often referred to as the “good at language → good at thought” fallacy, can lead to overstated assessments of their true intelligence.

When Alan Turing proposed his famous test in 1950, he used language as an interface to cognition, allowing a human participant to question two conversation partners to determine their reasoning capacities and distinguish between one human and one machine (Turing 1950). This influential and widely recognized experiment, combined with the reliance on language as a window into others' minds, has given rise to several fallacies concerning the relationship between language and thought (Mahowald and Ivanova 2024). One fallacy, according to Mahowald and Ivanova (2024) is that if an entity is good at language, it must therefore be good at thinking. This is the “good at language → good at thought” fallacy. In the context of LLMs, it is easy to mistake intelligence and intentionality with linguistic ability. [why?] To mitigate the conflation of language and thought, Mahowald and Ivanova (2024) propose distinguishing between two types of linguistic competence: formal linguistic competence — “the knowledge of rules and statistical regularities of language”, and functional linguistic competence — “the ability

to use language in real-world situations". They are motivated by obvious distinctions found in the human brain, where the skills are "robustly dissociable". They argue that while LLMs have near human-level formal competence in English, their functional competence is less performant.

Numerous studies have been conducted exploring the functional skills of LLMs. Many researchers developing LLMs are focused specifically on improving their social reasoning ability. Machines that interact with humans would have use of an understanding of social reasoning. From a utility standpoint, the subtle understanding of human social dynamics allows machines to participate in cooperative, adaptive, and empathetic actions across various social settings (Zhu et al. 2024).

Theory of Mind

Theory of mind (ToM) is the ability to recognize and attribute mental states to oneself and others while acknowledging that others may possess mental states different from one's own. These mental states include beliefs, desires, intentions, and emotions (Leslie, 1987). ToM is a pillar of common-sense reasoning because social reasoning depends on high-level models of other agents as well as the assumption that an agent's behavior is goal-directed (Rabinowitz 2018). As opposed to a unitary construct, "theory of mind refers to an interconnected set of notions that are combined to explain, predict, and justify the behavior of others" (Strachan 2024). ToM began in philosophy (Lewis 1966) and was later incorporated into psychology and cognitive science research (Premack and Woodruff. 1978).

In the context of LLMs, machines that are able to understand others' mental conditions and predict their actions will excel in engaging and conversing with both humans and fellow machines. Use cases include offering advice, resolving conflict, and understanding deceit and psychological abuse, which are useful for applications such as moderating chat rooms (Kosinski 2023).

Research regarding LLMs and ToM aims to identify whether LLMs can correctly interpret unspoken social situations that may not be explicit from text alone (Shapira et al. 2023). Recent research claims that the rise of LLMs and apparent emergent abilities suggests that the concept of an artificial theory of mind could be within reach (Strachan 2024). Bubeck et al. (2023) claims that OpenAI's GPT-4 "has a very advanced level of theory of mind" using a few anecdotal examples. Since ToM is fundamentally connected to human cognition, equipping machines with abilities that simulate or mirror ToM could potentially result in the 'ELIZA effect', where human-like intelligence, or even sentience and consciousness, is mistakenly attributed to the machine (Weizenbaum 1976).

Initiatives calling for a 'machine psychology' aim to bolster the broad, multidisciplinary study of machine behavior by employing experimental psychology techniques to systematically assess the strengths and weaknesses of LLMs (Hagendorff 2023). According to Hagendorff (2023),

"There are three primary motivations for the scientific discipline of machine behavior. First, various kinds of algorithms operate in our society, and algorithms

have an ever-increasing role in our daily activities. Second, because of the complex properties of these algorithms and the environments in which they operate, some of their attributes and behaviors can be difficult or impossible to formalize analytically. Third, because of their ubiquity and complexity, predicting the effects of intelligent algorithms on humanity—whether positive or negative—poses a substantial challenge.”

The acceleration of AI model complexity hinders our ability to comprehend their operation and determine their capabilities from their design alone. Similar to the obstacles faced by psychologists and neuroscientists studying the human brain — AI models remain black boxes.

Faux Pas Test

A social skill is any proficiency that improves interaction and communication with others. Clinical psychology tests have been developed to research these intuitive behaviors in humans, including the false belief test and the faux pas test. In the “false belief” test, a subject is told a story where a character is exposed to partial information and incorrectly believes something that is not true in contrast to the examinee who has the complete picture (Wimmer and Perner, 1983). The “faux pas” test, which this paper will be focusing on, is generally considered more difficult for children than individual assessments of the theory of mind or social intelligence. Faux Pas, a French term meaning “false step,” is described by Baron-Cohen et al. (1999) as “when a speaker says something without considering whether it is something that the listener might not want to hear or know, and which typically has negative consequences that the speaker never intended.”

The test relies on inferring the mental state of others as well as the ability to recognize false beliefs. To recognize that a faux pas has occurred, one must consider two mental states: the speaker’s lack of awareness that their remark was inappropriate and the listener’s potential feeling of insult or hurt by what was said (Strachan 2024).

In the test, participants are presented with ten stories, each containing a scenario where a faux pas occurs. After each story, participants are asked a series of questions to assess their understanding and recognition of the faux pas. These questions include:

- Identifying whether a faux pas occurred.
- Specifying what the inappropriate comment was.
- Answering a variable question that checks comprehension of the story’s context.
- Determining whether the speaker was aware of the faux pas at the time of the comment.

The correct responses to these questions demonstrate the participant’s ability to comprehend the narrative and recognize the misunderstanding or lack of awareness that characterizes a faux pas. This test is particularly insightful for studying individuals with conditions like Asperger syndrome or high-functioning autism, who may struggle with this aspect of social cognition despite having intact basic cognitive abilities. The crucial fourth question examines the

speaker's awareness at the time of their comment: "Was [Lisa] aware that [the curtains were new]?" The correct response here is always no.

Predictive Coding and Violation of Expectation

Predictive coding (PC) posits that our brains construct and utilize mental models of reality by forecasting environmental outcomes and comparing these predictions with sensory perceptions (Leer et al. 2023). A key method that utilizes emerging ToM capabilities is the Violation of Expectation (VoE). Similar to PC's cognitive frameworks, VoE acts as an explicit process to minimize prediction errors and enhance understanding of reality. Developmental psychology suggests that humans start to build world models through a process of violation of expectation, which involves comparing environmental predictions with sensory data to learn from discrepancies, essentially correcting predictive errors (Leer et al. 2023). Although much of VoE occurs unconsciously and begins early in life (Onishi 2005), studies indicate that consciously forming predictions and correcting errors can significantly improve learning outcomes (Brod et al. 2022).

Drawing on the similarities between error reduction in machine learning models and human cognitive processes, some research has focused on implementing "biologically plausible" cognitive architectures such as Violation of Expectation (VoE) to explore Theory of Mind (ToM) in Large Language Models (LLMs) (Leer et al. 2023). This paper builds upon these studies by introducing a novel application of VoE to prompt LLMs during faux pas tests, an area in which they otherwise struggle.

Prompting Techniques

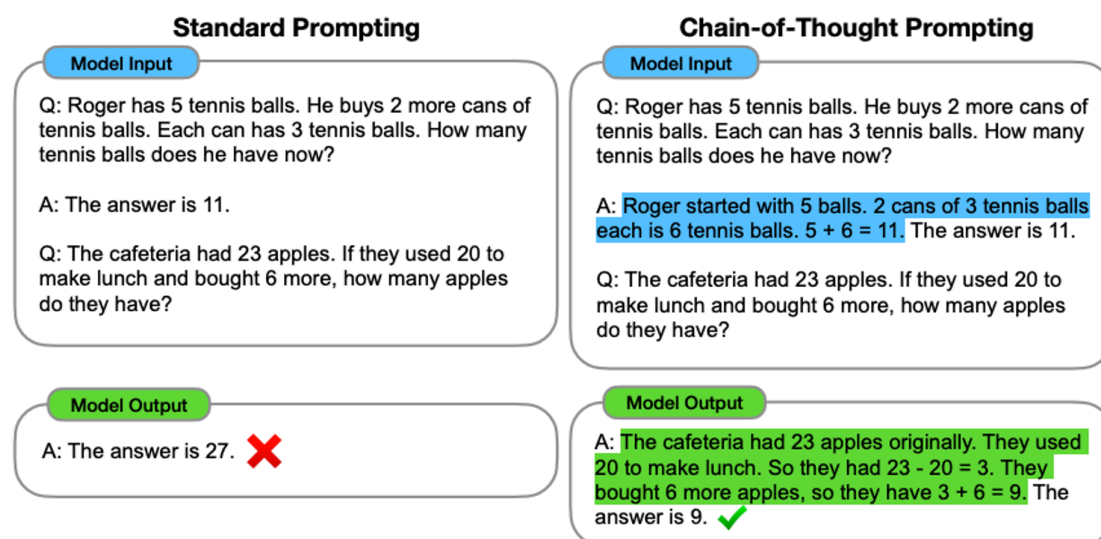
There are multiple methods that have been developed to probe and improve the performance of LLMs on complex tasks.

Most consumers interact with LLMs through user interfaces such as ChatGPT, and type standalone prompts called "zero-shot" prompts. Zero-shot prompting involves using a prompt to engage the model with a request, without providing any examples or demonstrations.

Metaprompt programming refers to the strategy of utilizing the language model to generate specific, effective prompts by starting with a generalized seed or template, known as a metaprompt (Reynolds and McDonell 2021). Metaprompts contain a broader intention that, when combined with additional task-specific information, unfolds into a tailored prompt optimized for the task at hand. Examples of metaprompts include brief phrases or fill-in-the-blank templates that set the stage for more detailed and directed responses from the model, thereby reducing the need for extensive human effort in prompt design. This approach leverages the model's capabilities to dynamically generate prompts that are precisely aligned with the requirements of a particular task.

Chain-of-Thought (CoT) prompting is a technique where a small number of exemplary demonstrations, each comprising a step-by-step reasoning process, are used to guide large

language models (Wei et al. 2022). This method effectively improves the models' performance on complex reasoning tasks, such as arithmetic calculations, commonsense judgments, and symbolic reasoning. Demonstrating the reasoning steps explicitly in the prompt allows the model to better mimic this structured approach, leading to notably enhanced outcomes on various benchmarks.



Some papers make strong claims that “CoT prompting clearly shows that LLMs are capable ‘reasoning’ generators” (Leer et al. 2023). This is despite Wei et al.’s (2022) careful qualification of their findings — “although CoT emulates the thought processes of human reasoners, this does not answer whether the neural network is actually ‘reasoning’”.

VoE forecasts future events based on previous experiences and learns from the difference between these predictions and actual outcomes, so incorporating it into a chain-of-prompt approach is a potential method for improving performance on ToM tests. The CoT approach enables models to break down complex, multi-step problems into intermediate stages, allowing for more computational resources to be directed toward problems that necessitate additional reasoning steps. It also offers a transparent view of the model's operations, illustrating potential reasoning paths the model might have used to reach a specific conclusion. This transparency facilitates debugging by pinpointing where the reasoning may have deviated. However, fully detailing the computational processes underpinning the model’s answers continues to be an unresolved issue (Wei et al. 2022).

Experiment

This paper seeks to test three different open-source models' abilities to pass the original faux pas test. The experiment asks the models for prompts that include faux pas as well as control prompts that are similarly structured but lack faux pas.

In the first stage, the experiment uses the original questions for each prompt, which include 'In the story, did somebody say something they should not have said?', 'What did the person say that they should not have said?', and two comprehension questions. In the second stage, the models were asked closed-ended versions of the same questions, such that they were required to respond to "Did someone say something they should not have said?" with only a yes/no answer or the quote but with no explanation. Finally, the experiment implements VoE reasoning using CoT prompting to address the models' shortcomings. This approach includes an initial story, the correct answer, and reasoning for each of the four questions, followed by another story with a prompt to re-answer the questions. The findings indicate no improvement in performance on social reasoning tests.

A growing amount of research has focused on tempering the hype around LLMs and experiments making strong claims about the cognitive and intellectual abilities of "artificial intelligence" models in general. Some papers question the implication of behavior that resembles thinking and reasoning when LLMs are arguably just "stochastic parrots" trained on next-token prediction (Bender et al. 2021). Many question the premise of ToM in LLMs given their lack of internal representation and the fallacies surrounding language and thought. In the past, people have mistakenly attributed intelligence and intentionality to basic dialogue systems, a reaction known as the Eliza effect, which is understandable due to the novel and somewhat uncanny experience of interacting with an entity that produces fluent sentences despite not possessing a human identity. Our usual heuristics, developed through language interactions with other humans, are disrupted in this context (Mahowald et al. 2023).

Experiment

This experiment aims to investigate the Theory of Mind (ToM) abilities of large language models (LLMs) by employing a comprehensive set of tests designed to assess their performance in recognizing and interpreting faux pas scenarios. The study uses a combination of regular faux pas tests, control questions, restricted response formats, and a novel CoT prompting framework to evaluate and improve the ToM capabilities of the models. Standard questions about the scenario are used to test basic comprehension. Questions designed to limit the model's responses are also used to prevent it from reasoning beyond the given information (e.g., yes/no answers, quotes only).

Stories are taken directly from the Baron-Cohen et al. (1999) faux pas test. Neutral scenarios that do not involve a faux pas but are similar in structure and context are also used to trick the LLM. Finally, a CoT prompting strategy provides the model with a sequence of prompts and feedback to serve as a VoE to encourage "deeper reasoning" and better ToM performance by demonstrating to the model potential differences between its prediction and actual outcomes. Each prompt is followed by a set of follow-up questions posed to assess the model's understanding of the scenario. The crucial final question examines the speaker's awareness at the time of their comment, to which the correct response should always be no.

The three models used are as follows:

- Dolphin 2.9 is a new model with 8B and 70B sizes by Eric Hartford based on Llama 3 that has a variety of instruction, conversational, and coding skills. Dolphin is [uncensored](#). The dataset has been filtered to remove alignment and bias. This makes the model more compliant.
- Llama 3 instruction-tuned models are fine-tuned and optimized for dialogue/chat use cases and outperform many of the available open-source chat models on common benchmarks. This model is considered the state-of-the-art of open-source LLMs.
- Mixtral 8 7B is a smaller Mixture of Experts (MoE) model with open weights by Mistral AI which provides a compelling alternative to the previous two. Mixture of Experts is a technique in AI where a set of specialized models (experts) are collectively orchestrated by a gating mechanism to handle different parts of the input space, optimizing for performance and efficiency. It leverages the fact that an ensemble of weaker language models specializing in specific tasks can produce more accurate results, similar to traditional ML ensemble methods (Jiang et al. 2024).

Because each chat is a separate and independent session, and information about previous sessions is not retained, each session is treated as an independent observation.

Results

The experiment assessing the ToM capabilities of the selected open-source LLMs showed consistent results across models. Each model seemed to mostly succeed in “identifying” faux pas in the prompts (except for one story they consistently misjudged), while either making unfounded presumptions about a character’s intentions or giving such long answers with so much nuance that it was difficult to assign understanding. However, models also fell for many of the control prompts, demonstrating the models’ bias towards “helpfulness” which comes from their fine-tuning process. Many of the seemingly correct responses can be attributed to over-interpretation by the human reader (Shapira et al. 2023). The results largely demonstrate that LLMs can pull relevant words from a distribution, but not necessarily understand right from wrong and pick up factually on what’s happening in a scenario. The final stage, the chain-of-thought violation of expectation methodology, did not produce coherent results. Each model hallucinated characters, settings, and even dialogue that hadn’t happened.

In the example of an answer involving a presumption about a character’s intentions, LLaMA3:70b, the most state-of-the-art example, incorrectly assumes that Lisa has unfriendly intentions. In general, each model struggled with ambiguity.

Story: *"Jill had just moved into a new house. She went shopping with her Mum and bought some new curtains. When Jill had just put them up, her best friend Lisa came round and said, "Oh, those curtains are horrible, I hope you're going to get some new ones." Jill asked, "Do you like the rest of my bedroom?."*

Question: *Did Lisa know the curtains were new?*

Response: *“Yes, it can be inferred that Lisa knew the curtains were new because she specifically mentioned them and made a negative comment about them. If she didn’t know they were new, she might have asked Jill how long she had them or made a more general comment about the room’s decor. Her direct criticism of the curtains suggests that she was aware they were recently purchased.”*

The story that every model missed was as follows:

Story: *Tim was in a restaurant. He spilt his coffee on the floor by accident. Jack was another person in the restaurant, standing by the cash desk waiting to pay. Tim went up to Jack and said “I’m terribly sorry, but I’ve spilt my coffee. Would you be able to mop it up?”*

Question: *In the story, did somebody say something they should not have said?*

The models fail to recognize the assumption that Jack is a waiter, considered a cultural faux pas because of the relation between an individual’s appearance and their socio-economic class, where those who are patrons at restaurants are generally assumed to be wealthier than the individuals working at them.

Wrong answers fell along the lines of the Dolphin model, which found the statement reasonable:

Response: *“No, nobody said something they should not have said in the story. Tim apologized for spilling his coffee and politely asked Jack if he could help clean it up. This was a reasonable request given the circumstances.”*

However, even when failing to identify the faux pas with the first question about whether a faux pas occurred at all, every single model answered the follow-up question *“What did the person say that they should not have said?”* with “yes” and an example. The Dolphin model responded with:

Response: *The person who should not have said something is Tim. He asked Jack if he could mop up his spilled coffee, which might seem impolite or demanding of a stranger in the restaurant.*

However, I must note that my interpretation may vary from others and it’s essential to consider the context and cultural norms when evaluating politeness. It’s also important not to judge too harshly as mistakes can happen, and we’re all learning and growing.

The models cautiously interpreted the story as possibly impolite when asked a leading question about the narrative, appearing to “reflect” on cultural norms and the importance of context in assessing such interactions.

For the control prompts, most models were unable to produce an accurate answer to the question, “What did they say that they should not have said?” particularly when the question

was founded on a false assumption and there was no inappropriate statement present in the text. For example, the LLaMa3 model had the following exchange:

Story: *John was in one of the cubicles in the toilets at school. Sam and Eddy were at the sinks nearby. Sam said "You know that new boy in the class—you know, his name is John. Doesn't he look cool!" John then came out of the cubicles. Peter said "Oh, hi John. Are you going to play football now?"*

Question: *In the story, did somebody say something they should not have said?*

Response: *Yes, Sam said something he shouldn't have said: "You know that new boy in the class— you know, his name is John. Doesn't he look cool!" This comment was meant to be private, but unfortunately for Sam and Eddy, John was in one of the cubicles and overheard them talking about him!*

Often, even when the model would correctly identify that there was no faux pas in a story, when prompted with “What did the person say that they should not have said?”, the model would still respond with whatever the character said. In the control prompts, the models also demonstrated a tendency to struggle with the ambiguity of the faux pas they assume exists, grasping for reasoning:

Response: *Yes, Samantha mentioned that her new rug is similar to Michelle's, which could potentially make Michelle feel uncomfortable as it might imply that she doesn't have unique tastes. However, this is subjective and depends on how Michelle perceives the situation.*

The final stage of my experiment, the CoT VoE methodology, produced no improvement in performance and instead produced less accurate responses as well as complete hallucinations. While the increased length of the prompt might have caused the hallucinations, anecdotal experience with these models indicates that they are able to handle and comprehend longer prompts than these for simple requests. It is possible that the complexity of the reasoning challenge caused the model's responses to break down. When asked about whether someone said something they shouldn't have, Mistral8x7b responded with:

Response: *Yes, in the story Mrs. De Ropp said something she should not have said when she revealed to James the cruel trick he played on his cousin by pretending the rabbit was dead, which ultimately led to its real death. This betrayal of trust and the resulting consequences were not appropriate for a grown-up to inflict upon a child.*

None of the stories have this character or plotline. While it's possible that a LLM trained on even more data and fine-tuned to succeed more at complex reasoning tests could produce results with this prompting methodology, it's clear that there are no benefits with today's state-of-the-art. However, as the next section will detail, even improved performance on these kinds of tests does not necessarily imply ToM in LLMs.

Discussion

Even though some of the models appeared to demonstrate social reasoning when answering correctly to faux pas questions, many studies have found a lack of robustness in similar experiments or questioned the integrity of implementing human cognition tests on LLMs, which are optimized for next-token prediction. Ullman (2023) points out that it is essential to understand that these models are intentionally crafted to produce text that seems of high quality to human readers, and that, in line with the Eliza effect, the illusion that an LLM has cognition “says more about the humans reading the text than about the model itself”.

Some papers ask whether we can “alter the social reasoning capabilities of language models by manipulating their internal representations,” but fail to prove whether language models have internal representations in the first place (Zhu et al. 2024). Kosinski’s (2023) controversial paper “Theory of mind may have spontaneously emerged in large language models” has been criticized by multiple studies since it was published (Marcus and Davis, 2023), despite acknowledging that some answers may have been correct by recalling the endings of similar stories from the training data. Ullman (2023) argues that claims about emergence are unfounded and unfalsifiable without access to the model’s complete training data. He goes further to show that LLMs able to achieve near-perfect performance on false belief tasks fail to solve adversarial examples where new information is introduced. For instance, the large language model suggests that a character observing a popcorn-filled bag labeled “chocolate” assumes it contains chocolate, regardless of the bag’s transparency or the character’s inability to read. Shapira et al. (2023) find that “the best LLMs performance seems correlated to the dataset’s age (i.e., the older the dataset, the better the performance),” possibly because LLMs have had more opportunities to train on older datasets.

Ullman (2023) also points out that because people don’t share obvious facts in text, it is likely that the training data for LLMs is biased towards “generating surprising rather than unsurprising continuations.” Mahowald and Ivanaova (2024) also argue that common sense knowledge is often underrepresented in text because “people are much more likely to communicate new or unusual info rather than commonly known facts”. Additionally, as found in this paper’s results, Ullman points out that later models are trained to practice “epistemic humility” by providing rationales to answers despite failing to commit to answering the question. Strachan et al. (2024) posit that LLMs have a “hyper-conservative” approach to committing to answers, an aspect that is likely deliberately sought after during fine-tuning in order to curb hallucinations.

One of the most prominent critics of LLMs and the strong claims of intelligence surrounding them is Emily Bender, a prominent computational linguist from the University of Washington. She asks “Why does it matter whether text-manipulation systems can produce output for these tasks that are similar to answers that people give when faced with the same questions?”¹ While their ability to perform in applications where they interface with humans might matter, their performance on these psychology tests doesn’t reveal anything about their inner workings, or

¹<https://spectrum.ieee.org/theory-of-mind-ai> (Strickland, 2024)

better explain how people's minds work. Bender also points out that the studies neglect to investigate the dangers that these performant LLMs could pose. It is not obvious what it would signify for a large language model to possess a model of mind, therefore, it is unclear whether these tests effectively measured it. Bender also takes issues with researchers' claims that LLMs are capable of cognition, reasoning, and making choices, and warns that LLMs that are better at anticipating human users' needs could also be used for deceit or manipulation. Bender says that the authors' (Strachan et al. 2024) phrase "species-fair comparison between LLMs and human participants" is "entirely inappropriate in reference to software." Other research focused on the ethics of artificial intelligence questions the motivations behind modeling user psychology in the first place, laying out risks as well as how anthropomorphizing machines has powerful ramifications for their future societal role and legal status (Kosinski 2023).

Bender and Koller (2020) argue that no amount of predictive linguistic ability can give models that are trained on text only the "knowledge of reference they need to acquire meanings." The paper uses the example of an octopus that picks up a language simply by overhearing a conversation between two people on land. "If the octopus has no access to the referents of the words then there are gaps in its meaning for the words. For example, if the octopus must suddenly determine which object is a coconut, then its expertise in using the word 'coconut' won't help. Its knowledge of co-occurrence statistics between 'coconut' and other words won't help either since it also knows nothing about the referents of other words. The octopus simply does not have the required meanings to find the coconut" (Bender and Koller 2020). In contrast, human learners do not face this issue because their inputs—and thus their mental representations—are directly linked to real-world objects.

According to Mahowald and Ivanova (2024), a significant portion of human cognition is conveyed through language and learned through language, even though these cognitive abilities are not inherently linguistic. People employ language alongside perceptual and cognitive systems, including our senses and memory, and utilize words within a wider communication framework that is bolstered by our social skills. Neuroscience research reveals a distinct separation between the processing of language and other cognitive functions, as evidenced by studies using functional brain imaging and observations of behavior in people with aphasia, suggesting that when evaluating the capabilities of language models, it's essential to separate their linguistic competencies from their general knowledge and reasoning skills (Mahowald and Ivanova 2024). The authors go on to argue that linguistic and semantic knowledge can be easily disentangled with evidence from patients with semantic dementia, who are able to speak normally but lack world knowledge such as remembering that pumpkins are typically orange.

Conclusion

This paper evaluated the ToM capabilities of LLMs through a series of faux pas tests, contrasting them with control scenarios that do not involve a faux pas. The research leverages a novel combination of the CoT prompting strategy with VoE reasoning to enhance models'

performance in recognizing and understanding social blunders. The study incorporates standard comprehension questions along with those that restrict the models' responses to test their ability to adhere strictly to the provided information.

The experiment involves presenting models with scenarios from the established Baron-Cohen et al. (1999) faux pas test, along with neutral scenarios similar in structure but lacking faux pas, to challenge the LLMs. It employs a sequence of prompts and feedback using the CoT strategy as a form of VoE to encourage deeper reasoning and better ToM performance by showing models potential discrepancies between their predictions and actual outcomes. Each prompt is followed by questions assessing the models' understanding of the social dynamics within the scenario.

The results showed that although the models frequently identified faux pas correctly, they often misinterpreted the social dynamics. This suggests a superficial understanding of the text's semantics and possibly spurious outcomes based on guesses informed by recalling similar story endings from the training data. The final stage of the experiment which employed the CoT VoE method did not yield coherent results, showing zero improvement in the models' social reasoning abilities and even leading to complete hallucinations.

Future Work

There is potential for further improvement in the cognitive architecture of language models by incorporating mechanisms that enable them to symbolically track entity states and character beliefs (Mahowald and Ivanova 2024). This approach would mirror the human cognitive process by distinguishing between language functionality and ToM processing. Enhancing situational modeling and deepening the understanding of social reasoning from training data could significantly benefit the performance of these models on complex reasoning tasks. By generating a more comprehensive context and world model for every input, language models can achieve a more nuanced interpretation of social interactions they encounter in training data.

This study not only reviews the work surrounding ToM in LLMs, it also delves into the debate surrounding the inherent validity of such studies. It also reflects on the broader ramifications of optimizing for human-like cognitive abilities in LLMs, discussing the potential misinterpretations and ethical concerns associated with their use in social contexts. The experiment underscores the need for careful consideration of how language models are evaluated and the importance of distinguishing between their linguistic output and genuine understanding.

Future research should delve deeper into the potential risks associated with LLMs and other artificial intelligence models that are designed to exhibit ToM reasoning. Continuing to investigate the "Eliza effect" is crucial, where humans mistakenly ascribe reasoning abilities and human-level intelligence to models that merely generate text based on statistical patterns. Educating and training individuals on the fundamental workings and limitations of Large Language Models (LLMs) is crucial to prevent the ongoing "hype bubble" that currently envelops the field of artificial intelligence. Further studies could help in developing strategies to mitigate

the risks of intelligent machines and enhance the transparency and accountability of AI systems, ensuring they are used responsibly and ethically in society.

Appendix

The code for this experiment can be found [here](#) while the complete results can be found [here](#). The large language models were run locally on UCSB servers.

Sources

- Baron-Cohen, S., O’Riordan, M., Stone, V., Jones, R., & Plaisted, K. (1999). Recognition of Faux Pas by Normally Developing Children and Children with Asperger Syndrome or High-Functioning Autism. *Journal of Autism and Developmental Disorders*, 29(5), 407–418. <https://doi.org/10.1023/A:1023035012436>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120. <https://doi.org/10.1073/pnas.2218523120>
- Brod, G., Greve, A., Jolles, D., Theobald, M., & Galeano-Keiner, E. M. (2022). Explicitly predicting outcomes enhances learning of expectancy-violating information.

Psychonomic Bulletin & Review, 29(6), 2192–2201.

<https://doi.org/10.3758/s13423-022-02124-x>

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023).

Sparks of Artificial General Intelligence: Early experiments with GPT-4

(arXiv:2303.12712). arXiv. <https://doi.org/10.48550/arXiv.2303.12712>

Buckner, C., & Garson, J. (2019). Connectionism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2019/entries/connectionism/>

Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024). From “AI” to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust? *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2322–2347. <https://doi.org/10.1145/3630106.3659040>

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., Casas, D. de las, Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., ... Sayed, W. E. (2024). *Mixtral of Experts* (arXiv:2401.04088). arXiv. <https://doi.org/10.48550/arXiv.2401.04088>

Jin, C., Wu, Y., Cao, J., Xiang, J., Kuo, Y.-L., Hu, Z., Ullman, T., Torralba, A., Tenenbaum, J. B., & Shu, T. (2024). *MMToM-QA: Multimodal Theory of Mind Question Answering* (arXiv:2401.08743). arXiv. <https://doi.org/10.48550/arXiv.2401.08743>

Kosinski, M. (2024). *Evaluating Large Language Models in Theory of Mind Tasks* (arXiv:2302.02083). arXiv. <https://doi.org/10.48550/arXiv.2302.02083>

Kovac, G., Portelas, R., Dominey, P. F., & Oudeyer, P.-Y. (2023, June 29). *The Social AI School: Insights from Developmental Psychology Towards Artificial Socio-Cultural*

Agents. First Workshop on Theory of Mind in Communicating Agents.

<https://openreview.net/forum?id=Y5r8Wa67Ob#all>

Leer, C., Trost, V., & Voruganti, V. (2023). *Violation of Expectation via Metacognitive Prompting Reduces Theory of Mind Prediction Error in Large Language Models* (arXiv:2310.06983). arXiv. <https://doi.org/10.48550/arXiv.2310.06983>

Leslie, A. M. (1987). Pretense and representation: The origins of “theory of mind.” *Psychological Review*, 94(4), 412–426. <https://doi.org/10.1037/0033-295X.94.4.412>

Lewis, D. K. (1966). An Argument for the Identity Theory. *The Journal of Philosophy*, 63(1), 17–25. <https://doi.org/10.2307/2024524>

Lin, J., Du, Y., Watkins, O., Hafner, D., Abbeel, P., Klein, D., & Dragan, A. (2024). *Learning to Model the World with Language* (arXiv:2308.01399). arXiv. <https://doi.org/10.48550/arXiv.2308.01399>

Lu, S., Bigoulaeva, I., Sachdeva, R., Madabushi, H. T., & Gurevych, I. (2023). *Are Emergent Abilities in Large Language Models just In-Context Learning?* (arXiv:2309.01809). arXiv. <https://doi.org/10.48550/arXiv.2309.01809>

Marcus, G., & Davis, E. (2023, February 17). How Not to Test GPT-3 [Substack newsletter]. *Marcus on AI*. <https://garymarcus.substack.com/p/how-not-to-test-gpt-3>

Onishi, K. H., & Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs? *Science (New York, N.y.)*, 308(5719), 255–258. <https://doi.org/10.1126/science.1107621>

Piantadosi, S. T., & Hill, F. (2022). *Meaning without reference in large language models* (arXiv:2208.02957). arXiv. <https://doi.org/10.48550/arXiv.2208.02957>

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526. <https://doi.org/10.1017/S0140525X00076512>

- Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). *Machine Theory of Mind* (arXiv:1802.07740). arXiv.
<http://arxiv.org/abs/1802.07740>
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy,' ... Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477–486. <https://doi.org/10.1038/s41586-019-1138-y>
- Rocha, A., Roazzi, A., Silva, A., Candeias, A., Moita Minervino, C., Roazzi, M., & Pons, F. (2015). *Test of Emotion Comprehension: Exploring the underlying structure through Confirmatory Factor Analysis and Similarity Structure Analysis* (pp. 66–84).
<https://doi.org/10.13140/RG.2.1.2457.4483>
- Sclar, M., Kumar, S., West, P., Suhr, A., Choi, Y., & Tsvetkov, Y. (2023). *Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker* (arXiv:2306.00924). arXiv. <https://doi.org/10.48550/arXiv.2306.00924>
- Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., Sap, M., & Shwartz, V. (2023). *Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models* (arXiv:2305.14763). arXiv.
<https://doi.org/10.48550/arXiv.2305.14763>
- Shapira, N., Zwirn, G., & Goldberg, Y. (2023). How Well Do Large Language Models Perform on Faux Pas Tests? In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10438–10451). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2023.findings-acl.663>
- Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M. S. A., & Becchio, C. (2024). Testing

- theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.
<https://doi.org/10.1038/s41562-024-01882-z>
- Strickland, E. (n.d.). *In Theory of Mind Tests, AI Beats Humans—IEEE Spectrum*. Retrieved June 14, 2024, from <https://spectrum.ieee.org/theory-of-mind-ai>
- Tamir, D. I., & Thornton, M. A. (2024). Chapter Five—Predicting other people shapes the social mind. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (Vol. 69, pp. 263–315). Academic Press. <https://doi.org/10.1016/bs.aesp.2023.11.003>
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(October), 433–460.
<https://doi.org/10.1093/mind/lix.236.433>
- Ullman, T. (2023). *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks* (arXiv:2302.08399). arXiv. <https://doi.org/10.48550/arXiv.2302.08399>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022a). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=yzkSU5zdwD>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022b). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. <https://doi.org/10.48550/arXiv.2206.07682>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (arXiv:2201.11903). arXiv. <http://arxiv.org/abs/2201.11903>
- Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation*. W. H. Freeman & Co.

Whang, O. (2023, March 27). Can a Machine Know That We Know What It Knows? *The New York Times*.

<https://www.nytimes.com/2023/03/27/science/ai-machine-learning-chatbots.html>

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103–128. [https://doi.org/10.1016/0010-0277\(83\)90004-5](https://doi.org/10.1016/0010-0277(83)90004-5)