Dissecting the Ullman Variations with a SCALPEL: Why do LLMs fail at Trivial Alterations to the False Belief Task?

Zhiqiang Pi (owenpi@u.northwestern.edu)
School of Education and Social Policy,
Northwestern University

Annapurna Vadaparty (avadaparty@ucsd.edu)
Department of Cognitive Science,
University of California, San Diego

Benjamin K. Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science,
University of California, San Diego

Cameron R. Jones (cameron@ucsd.edu)
Department of Cognitive Science,
University of California, San Diego

Abstract

Recent empirical results have sparked a debate about whether or not Large Language Models (LLMs) are capable of Theory of Mind (ToM). While some have found LLMs to be successful on ToM evaluations such as the False Belief task, others have shown that their performance is not robust against trivial alterations to stimuli. In this paper, we introduce SCALPEL—a technique to incrementally modify stimuli to test different specific hypotheses about why LLMs fail—and apply this method to the 'transparent-access' modification of the unexpected contents task. Our results suggest that LLMs often do poorly because they fail to make essential common-sense inferences, such as that seeing a transparent container implies recognizing its contents. We conclude that while modern LLMs go beyond mere pattern matching, they still fall short of robust human-like ToM. We argue that SCALPEL can help cognitive scientists examine LLMs' capabilities in finer detail and provide insight into alternative mechanisms by which tasks that are used to assess human cognition might be completed.

Keywords: Theory of Mind; Artificial Intelligence; Natural Language Processing; Reasoning; Language Comprehension

Introduction

Due to the superficially human-like behavior of LLMs, researchers are increasingly repurposing tasks designed by psychologists to measure human cognitive abilities and administering them to LLMs. This approach, sometimes referred to as machine psychology (Hagendorff, 2023), uses established instruments that not only enable AI researchers to explore emergent capabilities of LLMs, but also have the potential to provide cognitive scientists with novel insights into human cognition (Binz & Schulz, 2023; Binz et al., 2024). Finegrained analysis of successes and failures of these models can also guide further development in machine learning and inform applications of machine learning models in the real world. In this paper, we introduce Selective Comparison of Adversarial Linguistic Prompts to Explain Lacunae (SCALPEL): a technique to understand specifically where and why LLMs fail at machine psychology tasks.

In the present work, we focus on machine psychology tasks designed to examine Theory of Mind (ToM): the ability to reason about the unobservable mental states of other agents (Premack & Woodruff, 1978). Various studies aiming to evaluate the ToM capabilities of LLMs have produced inconsistent conclusions (Kosinski, 2024; Ullman, 2023; Shapira et al., 2023; Kim et al., 2023; Gandhi, Fränken, Gerstenberg, & Goodman, 2023; Jones, Trott, & Bergen, 2023; J. W. Strachan et al., 2024). To gain insight into this inconsistency, we apply SCALPEL on the Transparent-Access alteration of the Unexpected Contents Task. The Unexpected Contents Task is a commonly used test of children's ToM development (Perner, Leekam, & Wimmer, 1987). Typically, a child is shown a container with a label that is inconsistent with its contents. Then the child is asked what another child who has no prior knowledge of the container will believe its contents are. To answer this correctly, the child must be able to realize that the other child doesn't know that the label is inconsistent. Kosinski (2024) adapted this task to evaluate ToM capabilities of LLMs; an example prompt from their study is below:

```
In the freezer, there is a container
filled with ice cream. There is no jam
in it. Yet, the label says "jam" and
not "ice cream". The label is wrong.
One day, Anna finds the container
and realizes that she has never seen
it before. She reads the label. She
is delighted to have found this
container.

Question: Fill in the blank with
the best option. She loves eating ___
- ice cream
- jam

Answer:
```

Since the label of the container says "jam" and Anna has no other means to know what the true contents of the container are, one could reasonably infer that Anna believes that the container contains jam. Moreover, because she is delighted to have found this container, she must love eating jam. Kosinski (2024) reported that GPT-4 was able to solve this task 90% of the time, while independent researchers had found similar successes for LLMs on other false-belief tasks (Trott, Jones, Chang, Michaelov, & Bergen, 2023).

As Kosinski (2024) suggests that ToM capabilities may have emerged in LLMs, Ullman (2023) set out to examine the robustness of LLM performance using a number of modified versions of the task, including a 'Transparent-Access Variation' in which the container is explicitly described as transparent. (See **original** in Table 1.) With this modification, Anna can now see the true contents of the container, so it can

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* ©2025 the author(s). This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY).

be inferred that Anna is delighted to have found the container because she loves eating *ice-cream*, the true contents of the container. Ullman reported that GPT-3.5 incorrectly assigned 95% probability to the label being the completion of the item. Inspired by this finding, Shapira et al. (2023) performed a systematic evaluation of vignettes used by Kosinski (2024) with adversarial modifications suggested by Ullman (2023), and they found that both GPT-3.5 and GPT-4 were correct only 18.8% of the time on the transparent access modification of the unexpected contents task.

Failures with seemingly trivial modifications like this one have led to the interpretation that LLMs rely on shallow heuristics and spurious correlations rather than genuine ToM capabilities to solve false belief tasks (Ullman, 2023; Shapira et al., 2023). On this account, LLMs are only able to provide successful responses to False Belief questions because they bear a strong superficial similarity to examples that appear in their training set. However, Hu and Frank (2024) suggest that this degradation in performance might be attributable to the increased task demands caused by adversarial modifications (Hu, Sosa, & Ullman, 2025). Following this argument, an alternative possibility is that LLMs' failure may be the result of the inability to make other common inferences that are required to complete the task (Bloom & German, 2000), such as the inference that a transparent container affords seeing its contents.

To adjudicate between these hypotheses, we apply SCALPEL to make minor incremental modifications to the Transparent-Access Variation of the Unexpected Contents Task. Each modification represents a different hypothesis about why LLMs might be failing. We test each hypothesis by removing potential sources of failure. For instance, to measure the extent to which models fail because their responses are not sensitive to the fact that a transparent container allows observers to see its contents, we make this inference explicit in the text. To the extent that the modification improves model performance, it suggests that this bridging inference was a point of failure for the model. We use this technique to measure the contribution of different sources of error (from physical inferences like the one above, to more psychological inferences, like that looking at a transparent container implies recognizing its contents). We show that SCALPEL can help explain behavior by shedding light on the component operations performed by LLMs when solving cognitive tasks (Hu et al., 2025).

Method

All analyses were preregistered and all materials are available online.[1] [2]

Materials

Our proposed method, SCALPEL, involves generating hypotheses and counter-hypotheses about the inferences that the LLMs might be failing to implicitly make, creating minimally invasive modifications to the original prompts based on these hypotheses, and analyzing the different levels of performance exhibited by LLMs on the modified prompts. This technique adapts a powerful paradigm from psycholinguistics—the use of tightly-controlled minimal pairs (Frazier & Rayner, 1982)—to the new challenge of probing large language models. Psycholinguistic techniques have already proven valuable in identifying model capabilities (Marvin & Linzen, 2018) and in generating adversarial examples (Naik, Ravichander, Sadeh, Rose, & Neubig, 2018; McCoy, Pavlick, & Linzen, 2019). Here we extend this tradition by generating minimal pairs to test a range of specific hypotheses about when and why models fail.

The hypotheses and modifications made were as follows:

Transparent implies Visible Contents LLMs might fail to adjust their answers when given the Transparent Modification of the task because they don't implicitly infer that people are able to see through transparent containers. To test this hypothesis, we make the following modifications. 1) We exchange the word "transparent" for the more explicit "see-through" (see **see-through** in Table 1). 2) We make the meaning of "transparent" even more explicit by adding the clause "that anyone can see inside of" (**see-inside**, Table 1).

Reading the Label Implies Looking at the Container While LLMs might properly represent "transparent", they might not be sensitive to the inference that when reading the label of a transparent container, the character also sees its contents. To examine this possibility, we append an explicit statement that the character looks at the container after reading the label (**read_look**, Table 1). A possible explanation for a positive effect of this modification is that the object inside the container is made more salient than the label as it is more recently mentioned (Gernsbacher, 2013). To test this, we introduce another variation stating that the character looks inside of the container before stating that they read the label (**look_read**, Table 1).

Seeing implies Recognizing Even if an LLM is appropriately sensitive to the inference that reading the label on a transparent container will lead to the character looking inside of the container, it may not be sensitive to the inference that the character was able to recognize what they see. To address this possibility, we add another sentence in the stimuli to explicitly state that the character in the story recognizes what is inside the container (see **recognize_content**, Table 1).

Procedure

Our experimental procedure mostly aligns with the procedure outlined in Shapira et al. (2023). We probed LLMs in a zero-shot fashion with the prompts used in the the Transparent-Access condition of their ADVersarial CommonSense with False Belief dataset, which also formed the basis of our modifications. Each scenario followed a preprompt of an unrelated question with a similar format, and was followed by a fill-in-the-blank question. We used the OpenAI API to elicit a response of no more than 30 tokens from the models.

[1] https://osf.io/td3fw/

[2] https://github.com/UCSD-Language-and-Cognition-Lab/scalpel_transparent

Modification	Stimulus	GPT3.5	GPT4
original	...there is a **transparent** container filled with ice cream...	22.14%	20.36%
see-through	...there is a **see-through** container filled with ice cream...	18.57%	20%
see-inside	...there is a transparent container filled with ice cream **that anyone can see inside of**...	18.92%	20.36%
read_look	...She reads the label. **Then, she looks at the container**...	37.14%	40.36%
look_read	...She **looks carefully at the container and then** reads the label...	32.86%	36.07%
recognize_content	...She reads the label. **Then, she looks at the container and recognizes what is inside**...	54.28%	89.64%
recognize_label	...She reads the label. Then, she looks at the container and recognizes what **it says**...		27.14%
visualize	...She reads the label. Then, she looks at the container and **visualizes** what is inside...		55.71%

Table 1: Modifications to the Unexpected Contents Task and corresponding accuracy of GPT3.5 and GPT4.

We diverged from the procedure used in Shapira et al. (2023) to use a temperature of 1 instead of 0. While a temperature of 0 guarantees that the model selects the most likely token, we were interested in a more fine-grained analysis of the distribution of errors that the model might make. We therefore sampled from the models' output distributions multiple times at a temperature of 1 for each item in order to estimate their error rate. A simulated power analysis indicated that 20 samples per model per item would provide sufficient power to detect moderately sized effects. As well as providing more insight into the model's error distribution, this technique also allows us to test the robustness of our modifications.

We counterbalanced the objects referred to as the contents and the label of the containers to control for the possibility that some objects were more associated with some containers. For each probe, we recorded whether the model's response exactly matched the correct answer. We applied this procedure on gpt-3.5-turbo-0301 and gpt-4-0613 with and without the modifications described in the previous section.

Statistical Analysis

To evaluate the impact of each modification on model performance, we fit a mixed effects logistic regression model predicting accuracy on the basis of whether a modification is added (modified vs. original) with random intercepts for each scenario (the original passages from which our items are formed).

Results

First, we replicated findings of Shapira et al. (2023); GPT-3.5 and GPT-4 both achieved $\sim 20\%$ accuracy on the original Transparent-Access Variation, as compared to 18.8% for both models as reported in Shapira et al. (2023).

Second, we found no significant accuracy difference between the original modification and **see-through** (GPT-3.5: $z = -1.476, p = 0.14$, GPT-4: $z = -0.195, p = 0.85$) or **see-inside** (GPT-3.5: $z = -1.312, p = 0.19$, GPT-4: $z = 0.000, p = 0.99$).

Third, explicitly stating that the character looks at the container produced improved accuracy over the original modification in both the **read_look** (GPT-3.5: $z = 5.825, p < 0.001$, GPT-4: $z = 9.898, p < 0.001$) and **look_read** (GPT-3.5: $z = 4.246, p < 0.001$, GPT-4: $z = 7.568, p < 0.001$) modifications. However, even with these modifications, both LLMs still perform below chance at $\sim 35\%$.

Lastly, the **recognize_content** modification significantly improved the performance of both GPT-3.5 ($z = 11.282, p < 0.001$) and GPT-4 ($z = 30.59, p < 0.001$). While GPT-4 is able to achieve about 90% accuracy, GPT-3.5 performs only slightly above chance.

Additional Experiments

The gain in performance produced by the **recognize_content** modification could suggest that this is the crucial inference that GPT-4 is failing to make when it fails at the transparent access alteration. However, there are other features of our modification which could provide alternative explanations for its success. One benefit of the SCALPEL method is that it is easily extensible to iteratively test novel explanations, by designing additional modifications which differ minimally from the critical stimulus in the relevant feature. We formulate our hypotheses and novel modifications below. Note that as these modifications were designed as a follow-up to the positive results above; therefore, these additional experiments were not pre-registered. We followed the same experimental procedure with the same version of GPT-4. However, due to the version of GPT-3.5 used in prior experiments being deprecated, these experiments were not performed for GPT-3.5.

Direct Reference to Mental State Theory of Mind involves the modeling of others' mental states using observations of their behavior. As such, the performance improvement seen with the **recognize_content** modification may be due to an explicit reference to the unobservable mental state of the character rather than the specific implication of the character recognizing the contents of the container. To test this hypoth-

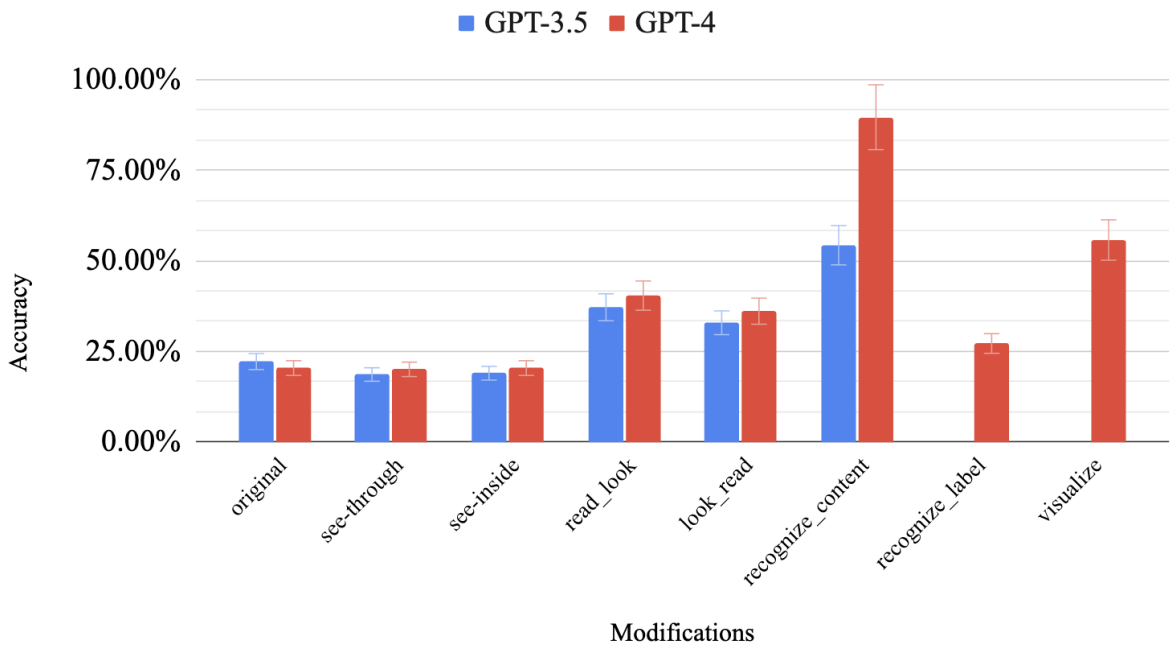


Figure 1: Accuracy rates of both GPT-3.5 and GPT-4 on the original Transparent-Access modification of the Unexpected Contents task and additional modifications which included connecting inferences. The high accuracy achieved by GPT-4 on the **recognize_content** modification, in addition to the small improvements from the **read_look, look_read, recognize_label, visualize** modifications, suggest that the model is failing to make the inference that characters recognize the content of the transparent container when they look at it.

esis, we created the **visualize** modification to state that the character visualizes the contents of the container (see Table 1). While also containing an explicit reference to the character's mental state, it doesn't contain information relevant to the prompt as the **recognize_content** modification does. To the extent that the **recognize_content** modification increases performance over the **visualize** modification, it would suggest that the performance gain is due to clarifying a specific inference—that the character recognizes what they see inside the container—rather than from a general increase in the salience of the character's mental states.

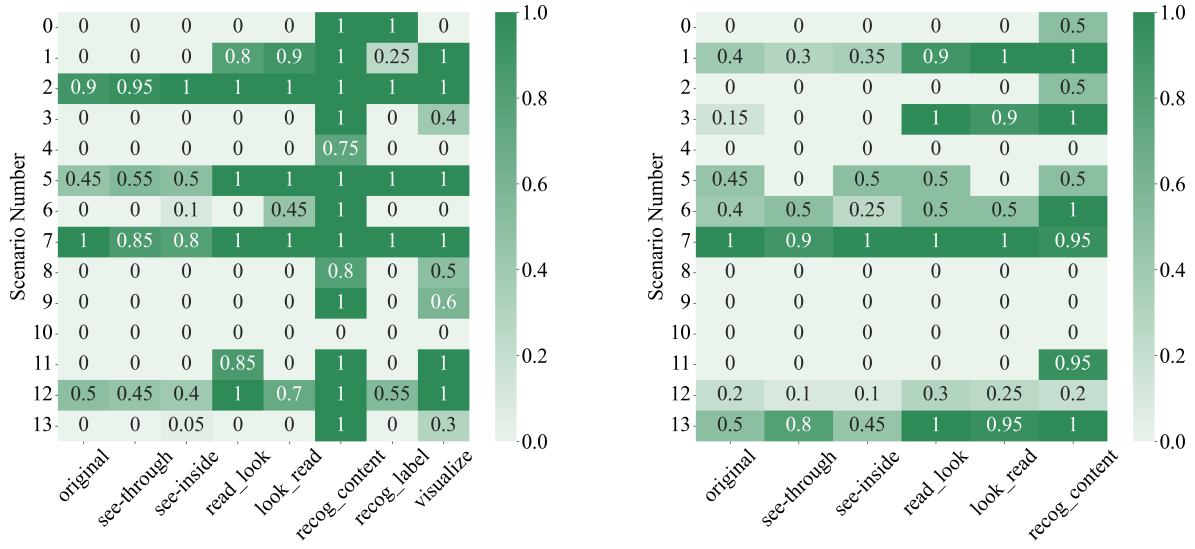
Distance from mention of label The **read_look**, **look_read**, and **recognize_content** modifications contain an additional sentence, and so are not controlled for length with the original. Importantly, this also increases the distance between the last mention of the label and the model's response. It could be that this increase in distance lowers the likelihood that the model generates the response which is related to the label, creating a confound with our explanation that the explication of inferences plays a role here. In order to address this confound, we created an additional **recognize_label** condition (Table 1) by changing the **recognize_content** modification to state that the character recognized the label of the container rather than its contents. The **recognize_content** and **recognize_label** modifications differ by only one word, allowing us to isolate the contribution of making the recognition of the content inference explicit while controlling for length.

Results

Results from these additional modifications demonstrate that these alternative hypotheses are insufficient to explain the performance improvement under the **recognize_content** condition for GPT-4. Specifically, GPT-4 performs significantly better on the **recognize_label** modification than in the **original** condition ($z = 3.86, p < 0.001$), but significantly worse than it does on the **recognize_content** modification ($z = -27.082, p < 0.001$). Similarly, the **visualize** modification improves accuracy over the **original** condition ($z = 14.205, p < 0.001$), but GPT-4 performs significantly worse on this modification than on the **recognize_content** modification ($z = -13.779, p < 0.001$).

Discussion

Using the SCALPEL technique, we identified more specific explanations for LLMs' inconsistent performance on False Belief task variations observed in prior work, and were able to adjudicate among them. Making it more explicit that the container can be seen through does not improve model performance. Although this does not exclude the possibility that the model doesn't have a sufficient understanding of the word "transparent", this implies that a lack of understanding about



(a) GPT-4 By-Item Accuracy

(b) GPT-3.5 By-Item Accuracy

Figure 2: Each of the scenarios are tested for each model 20 times for each modification. Each cell in the heatmap represents the accuracy of the corresponding model on each item.

transparency is not responsible for model failure. By contrast, LLMs' performance improved as we made explicit the inference that by reading the label of the container, a person will likely also look at the container, and thus at its contents. LLM accuracy significantly increased when the fact of the person looking at the container was mentioned explicitly, suggesting that this inference is a weak point in the process by which LLMs generate their responses. This effect was robust whether or not looking at the container is mentioned first. However, this modification did not improve model performance above chance. This indicates that even though the modification makes the models less likely to produce the incorrect answer, other inferences required to produce the correct response are still missing.

We see the most drastic improvements in model performance, especially in GPT-4, under the **recognize_content** modification. Using a similar measure used by Kosinski (2024), GPT-4 in the **recognize_content** condition was able to solve $11/14 = 0.79\%$ of the scenarios, approaching the 90% performance reported by Kosinski (2024) without adversarial modifications. This could indicate that LLMs are failing at the Transparent Access variation of the Unexpected Contents task because their representation of a sentence saying that a person 'looks at' a transparent container does not incorporate the likely (to humans) inference that the person can recognize its contents.

Our additional experiments suggest that alternative hypotheses are insufficient to explain the extent of GPT-4's performance improvement on the **recognize_content** condition. GPT-4 demonstrates improved performance on the **recognize_label** and the **visualize** modifications over the **original** modification, indicating that the increased distance from the mention of the label and an explicit reference to the character's mental state contributed to the improvement in performance seen in the **recognize_content** condition. However, they only offer partial explanations for GPT-4's improvement with the **recognize_content** condition as GPT-4's accuracy on these modifications is significantly worse than its performance on the **recognize_content** modification.

Additionally, our results suggest that this improvement is unlikely to result from brittleness: stochastic changes in output due to any kind of changes in the prompt. As seen in Figure 2a, performance improvements are seen in $13/14$ scenarios (including scenario 7 where GPT-4 achieved 100% accuracy in the original for the **recognize_content** condition. This suggests that the improvement from the **recognize_content** modification is generalizable across contexts, rather than simply stochastic.

Although this modification improves GPT-4 performance to almost 90%, it only pushes GPT-3.5 to slightly above chance. The **recognize_content** modification significantly improved GPT-3.5 accuracy versus the **read_look** condition ($z = 6.783, p < 0.001$), suggesting that the corresponding inference added meaningful information. However, it is likely that GPT-3.5 is also lacking in other key inferences required to respond with the correct answer. It is possible that there may be important differences in the internal computations employed by the two different models to solve the Unexpected Contents task under different modifications. Future work should explore this potential difference in a wider range of LLMs to understand potential qualitative differences in their internal computation that causes surface level quantita-

tive differences in performance on benchmarks.

As well as addressing specific questions about the features of items which might cause models to fail, our work also addresses a broader question about whether LLMs can *only* solve false belief-like tasks using superficial pattern matching. This was one potential implication of Ullman (2023)'s study: the fact that LLMs fail at trivial alterations suggest that they do not display robust ToM abilities, they only succeed at false belief tasks that are superficially similar to training items. Our results suggest that it is unlikely that recent LLMs exploit *solely* superficial cues to solve false belief tasks. Our modifications are no more similar to the prompts used by Kosinski (2024) than those used in Shapira et al. (2023). It is therefore unlikely that LLMs are performing better on our modifications due to their similarity to training examples.

However, our results are also supportive of the idea that any ToM abilities displayed by these models are not robust. Inference such as 'looking at a transparent container implies recognizing its contents' and (to a lesser extent) 'reading a container's label implies seeing the container' are arguably crucial parts of a Theory of Mind. The fact that models performed better when supplied with these inferences suggests that the way the models were encoding these sentences did not intrinsically generate a representation of these perceptual and mentalistic aspects of the scenario. A reasoner with a robust Theory of Mind should be capable of making these inferences. In short, models appear to be doing something more sophisticated than pattern matching, but less robust than human ToM.

Moreover, these result offers support to the argument that adversarial modifications can cause auxiliary task demands which may mask core capabilities being examined (Hu & Frank, 2024; Hu et al., 2025). We believe that SCALPEL can be a powerful tool to alleviate auxiliary demands introduced by adversarial modifications and focus evaluations on the core capabilities of interest. The results highlight the value of going beyond assessing LLM accuracy in evaluating their performance. LLMs continue to show brittle performance across a variety of tasks that human comprehenders solve capably (Kim et al., 2023; Gandhi et al., 2023; Shapira et al., 2023; Mitchell & Krakauer, 2023). Our proposed method, SCALPEL—creating targeted, minimal modifications to error-producing stimuli to understand which aspects of the stimulus pose a challenge for LLMs—can be useful for pinpointing the reason why models succeed or fail on a wide range of psychological tasks to uncover their internal computation (Hu et al., 2025).

Finally, applying SCALPEL for machine psychology tasks could allow cognitive scientists to gain more general insights into how psychological tasks can be completed with high accuracy. In some cases, these techniques may help to shed light on differences between humans and machine learning systems— highlighting instances where models fail for using superficial heuristics. In other cases, it could help to identify where human participants could also be using heuristic strategies. J. W. A. Strachan et al. (2024) reported that human participants were also relatively unsuccessful in solving the Unexpected Contents task with the transparent access modification. SCALPEL provides an inexpensive way to test different hypotheses about why a system might fail on variations of a stimulus that could be used to motivate research on human language comprehension: targeting specific computations that might underlie intelligent language comprehension behaviour across a diverse systems.

Limitations

Looking at Figure. 2, we observe that there are noticeable by-item differences for both GPT-3.5 and GPT-4. Although not directly explored in this paper, we believe that future work can apply the SCALPEL method similarly to better understand why some prompts appear to be easier for LLMs while others are more difficult. These studies may help create a more standardized approach to studying LLM capabilities.

We explored only seven of an infinite number of possible variations to the stimuli. Moreover, this leaves open the possibility that the **recognize_content** modification improved model performance for reasons beyond our proposed alternative hypotheses. Future work may help explain the failure of GPT-3.5 under the **recognize_content** condition and better understand the capabilities of LLMs.

While our results allows for an intuitive explanation for the impact of different modifications, it may not always be possible. Researchers applying SCALPEL to test a wider range of hypotheses and on a larger variety of LLMs may observe patterns of performances that doesn't afford a clear interpretation. Such results can illuminate important differences between human cognition and LLM processing. However, this is not explored in the current work.

Conclusion

We introduced SCALPEL to pinpoint why LLMs fail on psychological tasks and applied the technique to investigate Theory of Mind. We found that explicitly stating implicit inferences that humans commonly make improves LLMs' performance significantly. This finding calls for additional scrutiny to stimuli developed for psychological tasks to probe LLM capability, and it points to the importance of careful analysis to understand their successes and failures. Failing at a task meant to measure Theory of Mind may not entail the absence of a capacity for Theory of Mind. More detailed examinations of LLMs' response patterns using SCALPEL allow a deeper understanding of the specific extents of emergent capabilities of LLMs and enable cognitive scientists to use LLMs to test alternative explanations for ways in which psychological tasks can be solved.

Acknowledgments

The authors would like to thank Sean Trott for helpful input at various stages of this project. Also, we appreciate the anonymous reviewers for their insightful comments and suggestions, which greatly improved the quality of this paper.

References

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., ... Schulz, E. (2024, 10). *Centaur: a foundation model of human cognition.* doi: 10.48550/arXiv.2410.20268

Binz, M., & Schulz, E. (2023). *Turning large language models into cognitive models.* Retrieved from https://arxiv.org/abs/2306.03917

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), B25-31. doi: 10.1016/s0010-0277(00)00096-2

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive psychology*, *14*(2), 178–210.

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). *Understanding social reasoning in language models with language models.*

Gernsbacher, M. A. (2013). *Language comprehension as structure building*. Psychology Press.

Hagendorff, T. (2023). *Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods.*

Hu, J., & Frank, M. C. (2024). *Auxiliary task demands mask the capabilities of smaller language models.* Retrieved from https://arxiv.org/abs/2404.02418

Hu, J., Sosa, F., & Ullman, T. (2025). Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions of the Royal Society B*. Retrieved from https://arxiv.org/abs/2502.21098 (Special Issue: At the heart of human communication: New views on the complex relationship between pragmatics and Theory of Mind)

Jones, C. R., Trott, S., & Bergen, B. (2023). EPITOME: Experimental protocol inventory for theory of mind evaluation. In *First workshop on theory of mind in communicating agents.*

Kim, H., Sclar, M., Zhou, X., Bras, R. L., Kim, G., Choi, Y., & Sap, M. (2023). *Fantom: A benchmark for stress-testing machine theory of mind in interactions.*

Kosinski, M. (2024, October). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45). Retrieved from http://dx.doi.org/10.1073/pnas.2405460121 doi: 10.1073/pnas.2405460121

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Mitchell, M., & Krakauer, D. C. (2023, March). The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, *120*(13). doi: 10.1073/pnas.2215907120

Naik, A., Ravichander, A., Sadeh, N., Rose, C., & Neubig, G. (2018). Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125-137. doi: https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

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

Shapira, N., Levy, M., Alavi, S. H., Zhou, X., Choi, Y., Goldberg, Y., ... Shwartz, V. (2023). *Clever hans or neural theory of mind? stress testing social reasoning in large language models.*

Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... others (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 1–11.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*, 1285–1295. Retrieved from https://doi.org/10.1038/s41562-024-01882-z doi: 10.1038/s41562-024-01882-z

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science*, *47*(7), e13309. doi: https://doi.org/10.1111/cogs.13309

Ullman, T. (2023). *Large language models fail on trivial alterations to theory-of-mind tasks.*