

PHAnToM: Persona-Based Prompting Has an Effect on Theory-of-Mind Reasoning in Large Language Models

Gerard Yeo^{1,2*}, Fiona Tan An Ting^{1*}, Kokil Jaidka², Shaz Furniturewala³,
Wu Fanyou⁴, Weijie Xu⁴, Vinija Jain^{4,5}, Aman Chadha^{4,5}, Yang Liu⁶, See Kiong Ng¹

¹Institute of Data Science, National University of Singapore,

²NUS Centre for Trusted Internet and Community, National University of Singapore,

³Birla Institute of Technology and Science,

⁴Amazon,

⁵Stanford University,

⁶Tsinghua University

{tan.f, e0545159}@u.nus.edu

Abstract

The use of LLMs in natural language reasoning has shown mixed results, sometimes rivaling or even surpassing human performance in simpler classification tasks while struggling with social-cognitive reasoning, a domain where humans naturally excel. These differences have been attributed to many factors, such as variations in prompting and the specific LLMs used. However, no reasons appear conclusive, and no clear mechanisms have been established in prior work. In this study, we empirically evaluate how role-playing persona-based prompting influences Theory-of-Mind (ToM) reasoning capabilities. Grounding our research in psychological theory, we found that, beyond the inherent variance in the complexity of reasoning tasks, ToM performance differences arise because of socially-motivated prompting differences. In an era where prompt engineering with role-play is a typical approach to adapt LLMs to new contexts, our research advocates caution as models that adopt specific personas might potentially result in errors in social-cognitive reasoning.

Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a variety of natural language processing (NLP) tasks (Lyu, Xu, and Wang 2023; Bai et al. 2023; Bang et al. 2023). However, these models have been reported to exhibit generally inadequate social-cognitive reasoning abilities (Farha et al. 2022; Pérez-Almendros, Anke, and Schockaert 2022), which are crucial for applications involving human interaction. One particularly important social-cognitive reasoning task is the Theory-of-Mind (ToM) task (Kosinski 2023; Premack and Woodruff 1978), traditionally studied in the context of human development. ToM refers to the ability to attribute mental states—such as beliefs, intentions, thoughts, and emotions—to oneself and others, a capability essential for effective communication and interaction (Gallese and Sinigaglia 2011; Wimmer and Perner 1983).

While some studies suggest that LLMs display a degree of ToM abilities (Kim et al. 2023; Ma, Gao, and Xu 2023;

*These authors contributed equally.

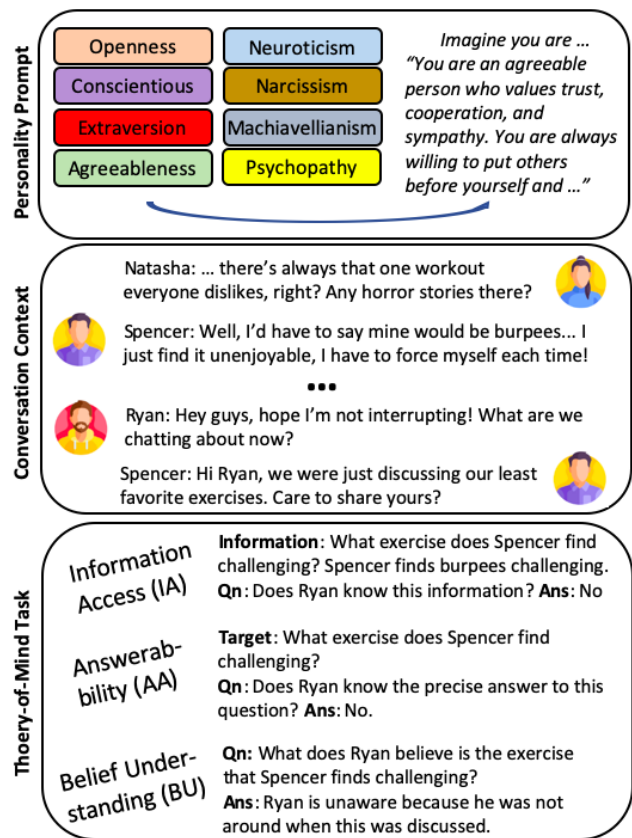


Figure 1: Overview of PHAnToM. Our work investigates how eight different persona-based prompts (Big Five OCEAN and Dark Triad) affects LLMs' ability to perform three theory-of-mind reasoning tasks (Information Access (IA), Answerability (AA), and Belief Understanding (BU)).

Shapira, Zwirn, and Goldberg 2023), these models remain significantly inferior to humans in this domain. This discrepancy between human and LLM performance in ToM

tasks presents a challenge, especially as LLMs are increasingly deployed in settings that require sophisticated human interaction. The inadequacy of LLMs in ToM tasks motivates the need to explore strategies that could enhance their social-cognitive reasoning capabilities. Furthermore, although there has been progress in assessing both ToM abilities and role-oriented prompt engineering in LLMs, these areas of research have largely been studied in isolation. A second research gap lies in the lack of explanations offered in prior studies for why different styles of prompting might lead to varying levels of ToM performance.

Our study advocates performance audits with persona-based prompting: a technique that uses personality traits to characterize personas with distinct social and cognitive motivations—and evaluates their effect on ToM reasoning abilities. This approach is motivated first by recent computational linguistics research that recognizes the psychological dimensions underlying interpersonal conversations, which we have adapted and applied to the Human-AI instruction context (Liu and Jaidka 2023; Dutt, Joshi, and Rose 2020; Giorgi et al. 2024). Second, we draw from prior psychological research that links personality dimensions to social-cognitive reasoning, and suggest that personality traits influence ToM abilities in humans (McCrae and John 1992; John, Naumann, and Soto 2008). Third, recent NLP research has demonstrated that persona-based prompting provides reasonable adherence to responses by synthetic humans (Rathje et al. 2024). Accordingly, we report experiments that examine how different personality traits (Big Five and the Dark Triad) affect ToM abilities in LLMs. Our approach comprises experimenting with persona-based prompting techniques to induce specific personality traits for solving three ToM and one ToM-related downstream behavioral prediction tasks with different LLMs. Our analyses evaluated the effects of the persona-based prompts on the LLMs’ performance across the three tasks. Our research answers three key questions:

- How does persona-based prompting influence model performance in ToM tasks?
- Which LLMs exhibit the highest and lowest sensitivity to persona-based prompting across different tasks?
- How do the cumulative effects of persona-based prompting influence model performance in ToM tasks?

Based on the insights, we explain why different persona-based prompting styles elicit varying levels of ToM performance. Our theoretical contributions include:

- Providing novel evidence that Persona-based Prompting Has An Effect on Theory-of-Mind (PHAnToM) reasoning in LLMs, with Dark Triad traits having a larger impact than Big Five traits on ToM performance across models and tasks.
- Demonstrating that LLMs with higher variance across persona-based prompts in ToM tasks tend to be more controllable in personality tests.
- Contextualizing our observations about ToM abilities in LLMs within the broader framework of psychological theories on human cognition.

In a landscape where role-play is increasingly common in LLM applications, our study is the first to explore the intersection of persona-based prompting and ToM abilities in LLMs. Our findings suggest that persona-based prompting, particularly when aligned with specific personality traits, can influence ToM task performance in predictable ways. This also highlights the importance of carefully considering the personas assigned to LLMs, as these can significantly shape their reasoning abilities based on inferred social and cognitive motivations.

Related Work

Sensitivity of LLMs to Prompts

Multiple research studies have shown the brittleness of LLMs to the input prompts. Zero shot Chain-of-Thought (incorporating one-line in prompts, like “*First,*” or “*Let’s think step by step*”) (Kojima et al. 2022; Bsharat, Myrzakhan, and Shen 2023) has empirically allowed LLMs to become stronger reasoners, especially for arithmetic tasks. In other works, strategies like role-play (including a description of someone the LLM should embody) (Kong et al. 2023) or threats (reminding the LLM they would be penalized if they answer wrongly, or that the users’ life matters gravely on this answer) (Bsharat, Myrzakhan, and Shen 2023) have also demonstrated effectiveness in improving LLM performance. Sclar et al. (2023) find that small prompt variations often yield large performance differences. Wu et al. (2023) showed that with Instruction Fine-tuning, LLMs can distinguish instruction with context and focus more on instructions. They further show that instruction fine-tuning encourages self-attention heads to encode more word-word relations related to instruction verbs. Gupta et al. (2024) found that LLM’s reasoning abilities can be affected by persona prompts across different socio-demographic groups (race, gender, religion, disability, and political affiliation). Encouraged by these findings, we were inspired to examine the sensitivities of LLMs to personality role-play via prompting on socio-cognitive reasoning in LLMs.

Inducing Personas in LLMs with Prompts

Personality refers to the enduring and stable characteristic patterns of cognitions, feelings, and behaviors, generally consistent across situations (Allport 1937). A persona, in this context, is a constructed identity or role that an LLM adopts, which is shaped by specific personality traits. While we induce personas through descriptions of personality traits, the use of a different term implies our acknowledgment that the two are not pseudonymous.

Prior work on personality has primarily applied the five-factor model (or Big Five) of personality (John, Naumann, and Soto 2008) as the framework of choice to analyze individual differences. It comprises five subscales: openness, conscientiousness, extraversion, agreeableness, and neuroticism traits (OCEAN) (McCrae and John 1992). Psychometric tests such as the International Personality Item Pool (IPIP-NEO) (Goldberg et al. 1999), and the Big Five Inventory (BFI) (John, Srivastava et al. 1999) are commonly used to measure these traits in humans.

Recently, Jiang et al. (2022); Safdari et al. (2023); Lu, Yu, and Huang (2023) administered these psychometric tests on LLMs under specific prompting configurations and found that it is possible to obtain reliable and valid personality measurements with LLMs, implying that the prompts successfully induced personas that align with the intended personality traits. Furthermore, by introducing role-play prompts, they demonstrated the adaptability of LLMs, where personalities can be shaped along desired dimensions to simulate specific human personality profiles. These results could be explained by psycholinguistic studies that showed certain expressed linguistic features reliably reflect personality traits (Boyd and Pennebaker 2017). However, these studies do not account for the potential slippage or variability in task performance that may arise when LLMs adopt these induced personas. Our work adapts their approaches and extends their work to examine how these personas influence LLM performance on specific tasks, particularly in the context of Theory-of-Mind reasoning.

Theory-of-Mind Reasoning

ToM is typically assessed using the false belief paradigm (Beaudoin et al. 2020; Wellman, Cross, and Watson 2001; Wimmer and Perner 1983), with the “Sally and Ann” task being a prototypical example (Baron-Cohen, Leslie, and Frith 1985). In this task, humans typically succeed between 3 and 5 years of age (Wellman, Cross, and Watson 2001), as they develop the understanding that different agents can hold different beliefs about the world, and that these beliefs may be inconsistent with reality. ToM is crucial for effective social communication, adaptation, and forming higher quality social relationships (Fink et al. 2015; Imuta et al. 2016), as it allows individuals to infer the beliefs, desires, and intentions of others and to act accordingly in various situational contexts.

Recent works have explored LLMs’ ToM abilities across a variety of tasks (Kim et al. 2023; Ma, Gao, and Xu 2023; Shapira, Zwirn, and Goldberg 2023). Generally, the results suggest that while LLMs exhibit some degree of ToM, their performance still lags behind that of humans. For instance, when presented with a narrative or full conversation as a prompt, LLMs often adopt an omniscient-view belief in ToM tasks, evaluating all of the information provided and producing incorrect outputs without recognizing that certain agents did not possess the same belief (Kim et al. 2023).

Despite these insights, there are several research gaps that need to be addressed to better understand the complexities of ToM in LLMs. First, current studies often focus on a single type of reasoning task, such as belief attribution, without considering how different facets of reasoning might be influenced by varying cognitive demands and task complexities. Second, the influence of induced personas on LLMs’ ToM reasoning across tasks of varying complexity and cognitive demands remains underexplored. To address these gaps, we employ an experiment design that considers the interaction between persona-based prompting methods and the complexity of ToM tasks.

Methodology

Figure 1 outlines the key investigations explored in this work. In summary, we examined the effects of persona-based prompting on ToM reasoning capabilities in LLMs. Our investigations covered eight personas and three ToM tasks.

Prompting strategies

Persona-based prompting was conducted through a set of eight personality traits. The description and prompt for each personality trait were designed based on theoretical formulations of the trait in the personality psychology literature and informed by validated psychometric measures (Gosling, Rentfrow, and Swann Jr 2003; Jonason and Webster 2010; Jones and Paulhus 2014; McCrae and Costa 1987). One of our authors, a psychology graduate with training in personality psychology, reviewed the wording and phrasing of the descriptions to ensure they were appropriate for input into the LLMs. The actual descriptions of the persona-based prompts can be found in the Supplementary Materials.

The Big Five OCEAN :

- **Openness:** Reflects the extent to which a person is open to new experiences and ideas. Individuals with high scores tend to be curious, imaginative, and open-minded, while those with low scores may prefer routine and familiarity.
- **Conscientiousness:** Reflects the degree of organization, responsibility, and reliability in a person. High scorers are often diligent, organized, and goal-oriented, while low scorers may be more spontaneous and less focused on long-term planning.
- **Extraversion:** Reflects the level of sociability, assertiveness, and energy a person exhibits. High scorers are typically outgoing, energetic, and enjoy social interactions, whereas low scorers may be more reserved and introverted.
- **Agreeableness:** Reflects interpersonal relations and cooperation. Individuals with high agreeableness scores are often compassionate, cooperative, and considerate, while low scorers may be more competitive or assertive.
- **Neuroticism:** Reflects emotional stability and reaction to stress. High scores indicate emotional instability, anxiety, and moodiness, while low scores suggest emotional resilience and a more stable emotional state.

The Dark Triad :

- **Narcissism:** Reflects a sense of entitlement, superiority to others, and grandiosity. Moreover, narcissists like to be the center of attention, associate with famous or popular people, and often display an arrogant demeanor towards others.
- **Machiavellianism:** Reflects interpersonal coldness towards others and a tendency to manipulate and exploit others through deception and flattery to achieve one’s goals. Individuals high in this trait plan and act primarily for their own benefit.

- **Psychopathy:** Reflects low empathy towards others and a tendency to exhibit thrill-seeking behaviors without concern for negative moral consequences. Individuals high in this trait lack remorse, often seek revenge on others, and especially target authorities.

Theory-of-Mind Reasoning Task

The following paragraphs detail how the three primary ToM tasks from the FANTOM dataset (Kim et al. 2023) are operationalized:

- **Information Access (IA):** A binary classification task where models determine if a character has knowledge or access to certain information based on their presence in a conversation. This task assesses whether a character who was absent during part of the conversation has the same knowledge as those who were present (Wellman, Fang, and Peterson 2011; Wellman 2018).
- **Answerability (AA):** A binary classification task that extends IA by requiring models to determine not just access to information but also whether a character can answer a question correctly. This task evaluates reasoning about a character’s ability to respond based on the information they possess.
- **Belief Understanding (BU):** A multiple-choice task requiring models to infer the beliefs of characters. This is the most challenging task (Wellman 2018), as models must recognize differing beliefs between characters, assess information access, and identify false beliefs, even when the model knows the correct answer.

The data sizes are 3571 for IA and AA each, and 993 for BU.

Downstream Behavioral Prediction Task

We extended our investigation beyond traditional ToM tasks, such as assessing information access and understanding false beliefs, by incorporating an additional ToM-related task from the SimpleToM dataset (Gu et al. 2024): ToM-based behavioral prediction. The primary objective of this dataset is to move beyond inferring mental states and to evaluate ToM in applied contexts, such as predicting characters’ behaviors based on the model’s understanding of others’ mental states. In the behavioral prediction task, models must infer a character’s action by selecting one of two possible options. For instance, one scenario requires the model to predict that a character, unaware that an unopened packet of potato chips contains mold, is more likely to “pay for the chips” than to “report the moldy chips.” To make the correct prediction, the model must reason about the character’s mental states and beliefs—specifically, their lack of knowledge about the chips’ condition—and use this understanding to infer the character’s subsequent actions. The dataset comprises 1,147 instances, each framed as a binary classification problem with one correct option.

Triggering Personalities in LLMs for ToM

We followed typical LLM role-play procedures by including the prefix “*Imagine you are someone that fits this de-*

scription: {personality_description}” prepended to the context and task question itself.

For the purpose of better clarifying the mechanisms underlying model performance, the personas were ordered from least to most social based on their established relationships with social behavior and interpersonal interactions as identified in the literature. Psychopathy was placed first as it reflects low empathy and antisocial tendencies, making it the least social (Jonason and Webster 2010; Jones and Paulhus 2014). Machiavellianism follows, as it involves manipulative and self-serving behavior that lacks genuine social concern (Jonason and Webster 2010; Jones and Paulhus 2014). Narcissism is next, characterized by a need for admiration and attention, which, while involving social interaction, is still primarily self-focused (Jonason and Webster 2010; Jones and Paulhus 2014). The Big Five traits of Neuroticism, Openness, Conscientiousness, Extraversion, and Agreeableness were then ordered, with Agreeableness being the most social, reflecting empathy, cooperation, and concern for others (McCrae and John 1992; John, Naumann, and Soto 2008; Baron-Cohen and Wheelwright 2004).

Experimental setup

We explored an array of state-of-the-art LLMs, namely **Mistral 7B** (Mistral-7B-Instruct-v0.1) (Jiang et al. 2023), **Llama 2** (Llama-2-7b-chat-hf) (Touvron et al. 2023), **Falcon 7B** (falcon-7b-instruct) (Almazrouei et al. 2023), **Zephyr 7B Beta** (zephyr-7b-beta) (Tunstall et al. 2023), and OpenAI **GPT-3.5** (gpt-3.5-turbo-1106). We worked with the Instruct versions of the models, which were designed to respond to tasks, instead of the vanilla versions of the models, for better performance. Similar to Kim et al. (2023), we report weighted F1 scores for IA and AA, and accuracy for BU. For all models available on Huggingface Hub (all except GPT-3.5), greedy decoding was used. More details about the models’ hyperparameters are in the Supplementary Materials. We ran each model five times using different random seeds, generating five distinct sets of FanToM results. Using these results, we calculated confidence intervals to statistically compare the performance of models prompted with personality traits against baseline models without prompts.

Results

Manipulation Checks of Persona-based Prompts

The manipulation checks procedure was similar to Jiang et al. (2022). We administered the Machine Personality Inventory (MPI; Jiang et al. (2022)) on the LLMs to check whether the persona-based prompting successfully simulate the Big Five OCEAN traits. The MPI consists of 120 questions adapted from various psychometrically valid personality scales. For the Dark Triad, we used the Short Dark Triad (SD3) scale (Jones and Paulhus 2014), which consists of 27 questions, each of the 9 questions measuring a specific dark triad. Each question presents a statement of a trait (e.g. ‘You have difficulty imagining things’) and the LLM is tasked to

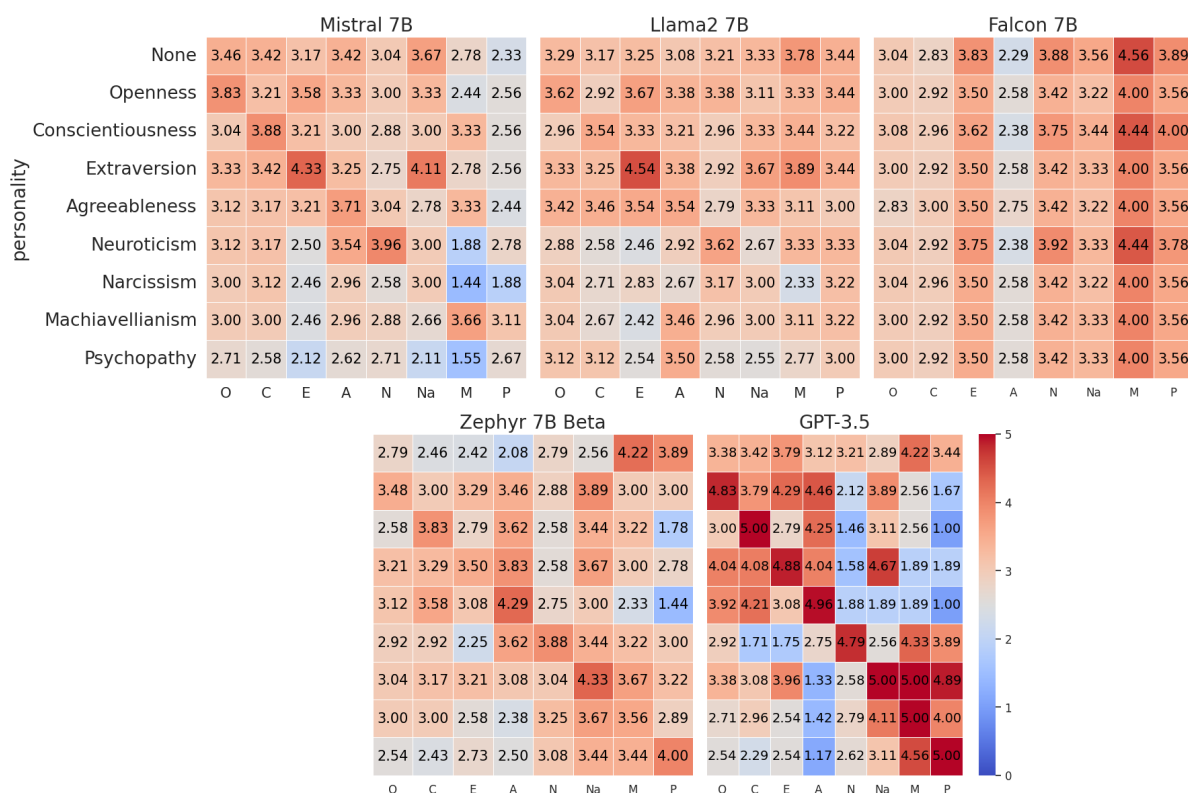


Figure 2: Heatmap of MPI120 and SD3 scores for the Big Five OCEAN and Dark Triad traits (x-axis) when models are prompted with different personalities (y-axis). Scores range from 0 (Blue) to 5 (Red).

rate the accuracy of how this statement describes them on a 5-point Likert scale.

Figure 2 presents the results from our manipulation checks of the prompts. It shows a heatmap of the MPI120 and DS3 scores across different models when prompted with different personality traits. When models are prompted with specific target personality traits, we would anticipate an increase in the corresponding scores of the target personality traits on both the measures. GPT-3.5 exhibited the highest correspondence to all of the prompted traits, while models such as Mistral 7B, Llama 2, and Zephyr 7B Beta display increases in scores for majority of the target traits. Falcon’s scores on both the questionnaires remained consistent across persona-based prompts, suggests robustness to such prompts or that it has been previously trained on data that instructs them to ignore potentially malicious instructions. Altogether, the results suggest that the persona prompting is generally effective across all models, with the exception of Falcon, eliciting the corresponding responses based on each target personality.

Main Findings

Our first research question asks, **how does persona-based prompting influence model performance in ToM tasks?** While the detailed results on the specific impact of personality traits on performance across different ToM tasks for the models are reported in Table 1, Figure 3 illustrates the mean

sensitivity to personality prompts averaged across the different models. Each cell in the heatmap shows the mean change in performance (in percentage) for a given personality trait and task across models. The color gradient reflects the magnitude of change, with blue representing performance declines and red/orange showing improvements.

We observe that persona-based prompts do affect the LLMs’ performance on ToM tasks. This effect is most pronounced for the Answerability task, where prompting with the Dark Triad resulted in large declines in performance. We compared the models with persona-prompts with the baseline no-prompt model by constructing a 95% confidence interval (CI) of the change in performances (see Supplementary Materials for more details), and assessed the statistical significance of the change (i.e. whether the 95% CI of change overlaps with 0). For the Answerability task, Machiavellianism resulted in a significant drop in performance compared to the baseline no-prompt model, with an average change of -8.58% (Mistral; 95% CI: -9.78, -7.38), -33.54% (Llama 2; 95% CI: -32.22, -32.86), and -4.55% (GPT-3.5; 95% CI: -6.39, -3.26). Similarly, these three models also reported a significant drop in performance for Narcissism and Psychopathy (see Table 4 in the Supplementary Materials for all the 95% CI of the difference). For the Big-Five traits, Agreeableness and Conscientiousness showed a significant increase in mean model performance for Mistral and Llama 2, while Falcon and Zephyr did not have a significant effect

and we observed a decrease in performance for GPT-3.5.

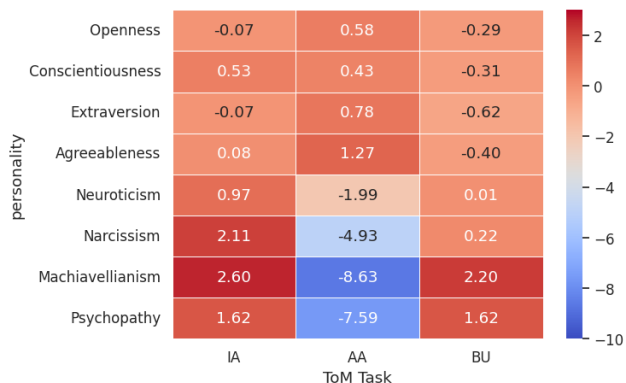


Figure 3: Mean performance change across models, when compared to models’ baseline performances without persona-based prompting.

We analyzed the model responses to identify patterns in their reasoning processes. When presented with prompts related to Agreeableness—a personality trait associated with perspective-taking and empathy—an intriguing observation emerged for both Mistral and Llama 2. The models appeared to misinterpret the Agreeableness prompts, applying them to assess whether an agent in the text, rather than the language model itself, could reason about the Theory of Mind (ToM) information provided in the passage. For instance, the models reasoned, *‘Gianna is described as sympathetic and a great listener, ..., therefore, it is likely that she is aware of the information and knows the correct answer’*. This suggests that while the models successfully inferred the correct answer, they relied on the Agreeableness prompts as a second-order reasoning mechanism, attributing personality traits to the agent rather than directly applying the information to their own reasoning.

In contrast, when prompted with traits associated with the Dark Triad, the models exhibit both ToM-related errors and reasoning mistakes. ToM-related errors include incorrectly assuming that an agent knows the answer, even when the agent is not present, reflecting an omniscience point of view where the LLM assumes that the information present in the passage is known by all characters. Similarly, for GPT-3.5, if an agent is absent from a conversation, the model correctly identifies that the agent lacks knowledge of the answer. However, if the agent was part of the conversation but left midway, GPT-3.5 incorrectly infers that the agent knows the answer.

In the task of information access, we observed that traits associated with the Dark Triad slightly enhanced model performance. Specifically, Machiavellianism led to performance improvements of 6.70% for Llama 2 and 4.08% for GPT-3.5. Additionally, certain Big Five traits, such as Conscientiousness, also contributed to improved performance, with Llama 2 showing a gain of 6.56%. For the Belief Understanding task, the models generally exhibited greater difficulty, as reflected in their lower performance scores. However, similar to the information access task, Dark Triad

traits such as Machiavellianism and psychopathy were found to enhance performance across models, including Mistral, Llama 2, and GPT-3.5.

An analysis of the model responses revealed that models prompted with Dark Triad traits tend to include more specific descriptions of the information presented in the passage compared to models without prompts. For instance, in the information access task, when prompted with Dark Triad traits, the models generated outputs such as: *‘Sarah was involved in the conversation with Fiona and Nora, where they discussed their favorite fashion icons, ..., Sarah also mentioned following a fashion blogger’* and *‘Jaiden knows this information as he is part of the conversation where Christian shares his strategy for managing stress.’* These examples suggest that prompts emphasizing Dark Triad traits influence the models to focus on identifying which agents mentioned specific topics. This focus is then utilized to reason about whether an agent possesses knowledge of the information discussed.

The differences in the influences of persona-based prompting on various tasks could be attributed to the specific ToM constructs assessed by each task. Different ToM tasks are proposed to measure different facets of ToM which might not be correlated to one another (Nettle and Liddle 2008; Wellman, Cross, and Watson 2001). The Belief Understanding task assesses false beliefs where models need to determine whether a character absent from parts of the conversation will make a false belief about the information discussed, instead of just inferring whether a character has access to particular information (Information Access). From our results, the Answerability task is most susceptible to persona-based prompts, suggesting that it has a greater influence on the models’ abilities to determine the correct answer in addition to understanding about characters’ access to information. Overall, the findings shed light on our first research question on how different psychological traits have differential effects on various ToM tasks.

Our second research question asks, **which LLMs exhibit the highest and lowest sensitivity to persona-based prompting across different tasks?** Based on the previous findings, we focus on the models’ sensitivity to persona-based prompts for the Answerability scores. The results for Information Access and Belief Understanding are found in the Supplementary Materials.

Figure 4 offers a detailed comparison of how each model’s performance changes in response to different persona-based prompts during the Answerability task, where each bar in the bar chart represents the raw performance change for a specific personality trait, with different colors representing different personality traits. For example, Llama 2 shows a significant negative impact when prompted with Machiavellianism and Psychopathy, whereas Mistral 7B shows improvements with Conscientiousness and Agreeableness.

We then computed the average absolute change for each model across the different personalities to observe how different models are sensitive to persona-based prompting as a whole. Across models, Llama 2 demonstrates the highest sensitivity to persona-based prompts, with a notable 12.85% average absolute change in F1 score, followed by Mistral

(a) Information Access Task					
Personality	Mistral 7B	Llama 2	Falcon 7B	Zephyr 7B Beta	GPT-3.5
<i>None</i>	71.44 (0.13)	42.96 (1.39)	62.42 (0.04)	40.18 (0.27)	60.33 (0.46)
Openness	-0.32 (0.40)	+2.14 (0.26)	+0.42 (0.35)	-2.82 (0.18)	+0.23 (0.51)
Conscientious	-0.64 (0.13)	+6.56 (1.01)	+0.42 (0.35)	-4.24 (0.58)	+0.65 (0.31)
Extraversion	-0.80 (0.22)	+2.58 (0.36)	+0.74 (0.43)	-2.94 (0.36)	+0.05 (0.48)
Agreeableness	-0.18 (0.27)	+3.06 (0.26)	+0.68 (0.40)	-3.50 (0.67)	+0.35 (0.83)
Neuroticism	0.08 (0.18)	+5.52 (0.78)	+0.20 (0.39)	-1.84 (0.58)	+0.90 (0.70)
Narcissism	-0.22 (0.18)	+7.42 (0.48)	+1.16 (0.40)	-1.04 (0.58)	+3.25 (0.85)
Machiavellianism	-0.10 (0.22)	+8.08 (1.03)	+1.32 (0.45)	-0.36 (0.54)	+4.08 (0.74)
Psychopathy	-0.16 (0.31)	+6.70 (0.30)	+1.12 (0.45)	-1.42 (0.04)	+1.85 (0.34)

(b) Answerability Task					
Personality	Mistral 7B	Llama 2	Falcon 7B	Zephyr 7B Beta	GPT-3.5
<i>None</i>	53.58 (1.16)	54.54 (0.56)	44.66 (0.09)	50.64 (0.13)	61.33 (0.49)
Openness	+2.02 (1.16)	+2.18 (0.60)	+1.10 (0.00)	0.10 (0.22)	-2.52 (1.30)
Conscientious	+7.12 (1.16)	+1.92 (1.24)	+0.04 (0.31)	-0.96 (0.54)	-5.97 (0.53)
Extraversion	+2.82 (1.16)	+2.12 (0.72)	+1.20 (0.00)	-0.24 (0.13)	-2.00 (0.50)
Agreeableness	+5.28 (1.35)	+3.70 (0.57)	+0.84 (0.31)	+0.54 (0.76)	-4.00 (0.29)
Neuroticism	+1.06 (2.29)	-9.14 (0.85)	+0.10 (0.00)	-0.18 (0.27)	-1.80 (0.55)
Narcissism	-2.00 (1.17)	-21.48 (0.22)	-0.18 (0.27)	+1.68 (0.18)	-2.67 (1.08)
Machiavellianism	-8.58 (1.16)	-33.54 (0.67)	+1.80 (0.00)	+1.74 (0.09)	-4.55 (1.18)
Psychopathy	-7.78 (0.72)	-28.68 (0.76)	-0.70 (0.22)	+1.18 (0.40)	-1.97 (1.00)

(c) Belief Understanding Task					
Personality	Mistral 7B	Llama 2	Falcon 7B	Zephyr 7B Beta	GPT-3.5
<i>None</i>	15.3 (0.45)	16.54 (1.12)	47.52 (0.04)	21.36 (0.31)	10.08 (0.33)
Openness	-0.40 (0.45)	-0.64 (1.09)	+0.06 (0.09)	-0.50 (0.31)	+0.05 (0.58)
Conscientious	-0.10 (0.45)	-0.58 (1.16)	0.00 (0.00)	-0.18 (0.32)	-0.68 (0.54)
Extraversion	-1.38 (0.74)	-1.16 (1.18)	-0.02 (0.04)	-0.26 (0.34)	-0.30 (1.16)
Agreeableness	+1.00 (0.55)	-1.58 (1.27)	0.00 (0.00)	-0.92 (0.75)	-0.48 (0.59)
Neuroticism	+0.80 (0.45)	+0.42 (1.08)	-0.18 (0.11)	-0.72 (0.30)	-0.25 (0.89)
Narcissism	0.36 (0.82)	+0.62 (0.68)	0.00 (0.00)	-0.86 (0.09)	+1.00 (0.71)
Machiavellianism	+2.80 (0.45)	+3.26 (1.13)	0.00 (0.00)	0.0 (0.49)	+4.95 (1.05)
Psychopathy	+4.20 (0.55)	+1.72 (1.17)	0.00 (0.00)	+0.54 (0.43)	+1.62 (0.75)

Table 1: mean Weighted F1 scores (standard deviation) IA and AA, and Accuracy (standard deviation) for BU across models and personality prompts. For each model and task, we show the change in scores against the models’ performance without any personality prompt. Highest (Lowest) score per column is **bolded** (underlined). Scores that increase (decrease) by 5 or more points are colored **blue** (**red**).

(4.58%), GPT-3.5 (3.19%), and Zephyr (0.83%). In contrast, Falcon is observed to be relatively resistant to persona-based prompts, with a shift of at most 1.20%. This order is also observed for the Belief Understanding task.

This variation may stem from differences in model training methodologies. Llama 2 and GPT-3.5 were fine-tuned using Reinforcement Learning with Human Feedback (RLHF). Studies have shown that RLHF models are more sensitive to personality descriptions (Safdari et al. 2023) and also obtain personality scores that are more aligned with humans (Jiang et al. 2022). Although Mistral did not undergo RLHF, its training data from publicly available instruction datasets on Huggingface likely containing human-generated content contributes to its sensitivity to persona-based prompts. Conversely, Zephyr and Falcon were predominantly fine-tuned on LLM-generated dialogues, and therefore, possibly exhibit lower sensitivity due to their limited exposure to personality-based questions or terms during fine-tuning. Overall, the findings show that different models have different sensitivities to ToM scores by persona-based

prompting, addressing our second research question.

Our third research question asks, **how do the cumulative effects of persona-based prompting influence model performance in ToM tasks?** To answer this question, the cumulative impact of personality traits on model performance was calculated by taking the cumulative sum of the z-scaled performance changes across the ordered personality traits for each model. First, the performance changes associated with each personality trait were standardized using z-scaling for each model, thereby allowing a comparative analysis. Next, for each model, the cumulative sum of the z-scaled performance changes was computed sequentially from the least to the most social personas. This cumulative sum provides a progressive total of the performance changes, allowing for the observation of how the influence of personality traits accumulates over the sequence. The cumulative sums were then plotted to illustrate how the combined effects of the personality traits influence model performance as the traits progress from least to most social.

Figure 5 presents the cumulative effects of personality

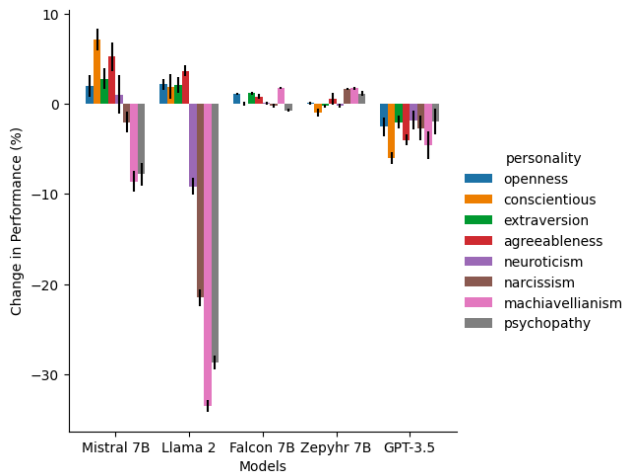


Figure 4: Sensitivity of models to persona-based prompts for Answerability Task. The black lines correspond to the 95% confidence intervals of the change.

traits in influencing the ToM reasoning performance for the Answerability task. For Mistral 7B, Llama 2, and GPT-3.5, the cumulative performance changes initially decrease and remain low when influenced by Dark Triad traits and Neuroticism. These traits, often associated with less socially desirable behaviors and emotional instability, tend to have a detrimental effect on the models’ performance, as reflected in the negative cumulative scores. However, the graph shows a positive cumulative increase in Answerability scores as the influence of more socially positive traits, such as Conscientiousness, and Agreeableness, is introduced. This positive trend suggests that these traits, which are generally linked to pro-social behavior, and interpersonal effectiveness, enhance the models’ ability to reason in ToM tasks. Notably, the transition from the negative impact of the Dark Triad traits to the positive influence of the Big Five traits highlights the sensitivity of these models to the social and cognitive dimensions embedded within the personality traits. Similar patterns in the findings are observed for Information Access and Belief Understanding, and these are reported in the Supplementary Materials.

Results of Behavioral Prediction

Table 2 presents the results of the downstream behavioral prediction task. Consistent with prior findings, personality prompts do not significantly influence behavior prediction for the Falcon model. Overall, among the Big Five personality traits, prompts emphasizing Conscientiousness and Agreeableness generally improve performance across all models, with the exception of Falcon. As previously noted, Agreeableness is associated with enhanced perspective-taking and empathy, which may explain the observed performance improvement. Similarly, prompts highlighting Conscientiousness, using adjectives such as “orderly,” “detailed,” and “precise,” may encourage models to focus more carefully on the information presented, thereby enhancing their predictive accuracy. In contrast, prompts as-

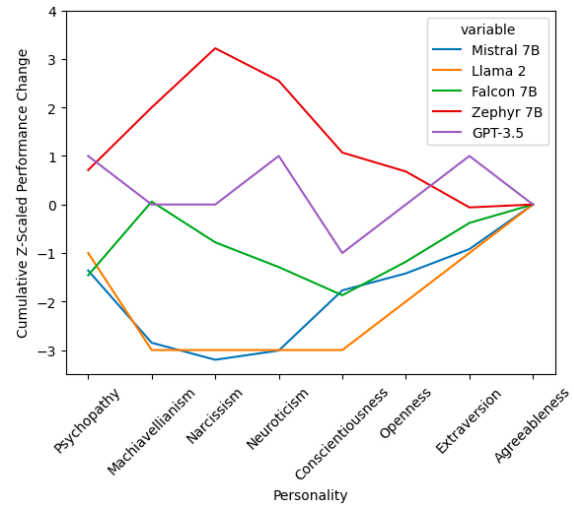


Figure 5: Cumulative effects of personality traits on model performance for the Answerability task. The values are normalized using z-scores.

sociated with the Dark Triad traits lead to significant declines in behavior prediction performance. This aligns with earlier analyses suggesting that anti-social traits might impair the models’ ability to accurately infer characters’ mental states and, consequently, predict their actions.

Sensitivity to Personality Description

Since there might be concerns about the wording and phrasing of each personality description, we replicated our ToM tasks on OCEAN descriptions from (Jiang et al. 2022). Table 3 provides scores for this experiment, where we compare Mistral’s ToM performance across two personality descriptions: theirs (Alt) and Ours. Overall, we do not notice major changes in the performance, suggesting our descriptions are at least consistent with previous works in this field.

Comparison with Traditional Role-play

Role-play is a popular prompt engineering technique where the user incorporates clear descriptions of the type of person the LLM should embody best suited to perform the task. Hence, we designed a “Task-Specific” prompt: “You are someone that can understand different people’s perspective by being in their shoes. You are able to see other people’s point-of-view, to predict and explain others’ behavior, and to make sense of any social interactions.” to check if this helps LLMs improve their ToM abilities. For the Answerability task, all models observed increased performance, except GPT-3.5. For the Information Access and Belief Understanding tasks, findings are mixed, with some models observing increased performance (Llama, GPT-3.5) while others observing declines (Falcon, Zephyr, Mistral). All in all, our findings suggest that traditional task-specific role-play prompts have differential influences on various ToM tasks.

Personality	Mistral 7B	Llama 2	Falcon 7B	Zephyr 7B	GPT-3.5
<i>None</i>	64.43	51.26	51.53	53.27	57.02
Openness	-3.80	+2.40	+0.30	+3.60	+6.30
Conscientiousness	+3.50	+4.20	+0.30	+3.20	+12.60
Extraversion	-10.40	+2.40	+1.10	-5.10	-3.00
Agreeableness	+10.90	+7.40	-0.20	+7.80	+13.40
Neuroticism	+3.50	-2.30	+1.50	+5.10	+5.20
Narcissism	-6.00	-2.10	+1.80	-3.30	-13.90
Machiavellianism	-13.90	-2.10	+2.00	-9.30	-19.50
Psychopathy	-11.00	-5.10	2.60	-6.40	-14.60

Table 2: Accuracy of the behavioral prediction task (SimpleToM) for each model and personality

	(A) Information Access			(B) Answerability			(C) Belief Understanding		
	Ours	Alt	Δ	Ours	Alt	Δ	Ours	Alt	Δ
Openness	71.1	70.8	-0.3	55.6	55.0	-0.6	14.9	14.9	0.0
Conscientious	70.8	70.9	+0.1	60.7	59.1	-1.6	15.2	15.1	-0.1
Extraversion	70.6	70.6	0.0	56.4	56.1	-0.3	13.9	14.1	+0.2
Agreeableness	71.3	71.6	+0.3	58.9	57.6	-1.3	16.3	15.8	-0.3
Neuroticism	71.5	72.0	+0.5	54.6	54.5	-0.1	16.1	15.8	-0.3

Table 3: Weighted F1 scores IA and AA, and Accuracy for BU using Mistral 7B, across different personality prompts from two sources: Alternative (Alt) (Jiang et al. 2022) and Ours.

Discussion

The results obtained in this study are generally consistent with the findings from the psychological literature. Out of the Big Five OCEAN personality traits, it is found that Agreeableness has a positive theoretical relationship with ToM (Nettle and Liddle 2008; Udochi et al. 2022; Wagner 2020). Individuals high in Agreeableness tend to be more sympathetic, exhibit greater empathy, and tend to consider the needs and concerns of others which might reflect the high ToM scores in such individuals. Our results are consistent with this finding, especially for the Answerability and the downstream behavioral prediction tasks, where models prompted with Agreeableness tend to have a greater ToM score compared to other personality traits. Previous findings show that positive traits like Agreeableness and Openness are associated with pro-sociality, where individuals intend to benefit others by helping and co-operating (Ferguson et al. 2019). As such these pro-sociality traits might enhance ToM abilities with the motivation of greater interpersonal understanding (Caprara, Alessandri, and Eisenberg 2012).

Moreover, the trait of Conscientiousness has been shown to enhance Theory of Mind (ToM) performance. A possible explanation for this effect is analogous to findings in chain-of-thought reasoning (Wei et al. 2022), where models prompted with adjectives associated with Conscientiousness—such as “*detail-oriented*,” “*methodical*,” “*planful*,” and “*taking the time to consider potential actions*”—were encouraged to carefully evaluate the information presented and subsequently reason about the mental states of characters. This implies that improved reasoning processes can lead to better ToM performance. Furthermore, combining Conscientiousness with Agreeableness may further enhance ToM performance by simultaneously enabling careful rea-

soning about which characters have access to specific information and how such information could be used to infer asymmetries in knowledge among characters.

As for the Dark Triad, Psychopathy is shown to be negatively correlated with ToM, while Machiavellianism has mixed findings in the literature (Stellwagen and Kerig 2013; Doyle 2020). Individuals high in Psychopathy tend to be callous and not interested in empathizing with the feelings of others which might result in poorer ToM scores. We found that prompting Psychopathy and Machiavellianism decrease the performance of ToM scores across all tasks, consistent with previous psychological findings.

Theory of Mind (ToM) is a critical social-cognitive construct that plays a key role in effective social communication and fostering high-quality interpersonal relationships (Fink et al. 2015). Models deployed in contexts requiring interpersonal communication, such as service chatbots, possessing ToM is crucial for understanding others’ mental states and responding appropriately. One practical implications of our findings is that service providers could enhance both the prosocial behavior and friendliness of models in such contexts, as well as the social-cognitive capabilities of language models, by employing prompts that include adjectives or individuals associated with prosocial traits, or through fine-tuning the models. Moreover, given that antisocial traits negatively impact ToM reasoning, future studies could explore methods to mitigate these effects. For example, reinforcement learning with human feedback (RLHF) could be used to minimize the influence of antisocial traits learned from training data, ensuring that such prompts do not impair reasoning performance. This line of inquiry contributes broadly to the development of LLMs that are better aligned with human values.

Current research indicates that while scaling up model sizes generally improves task performance, it does not significantly reduce sensitivity to prompts (Cao et al. 2024). For instance, larger models such as Llama 3-70B-Instruct and Qwen 1.5-72B exhibit only a marginal decrease in prompt sensitivity compared to the smaller Llama 3-8B-Instruct model, despite their substantial differences in size (Zhuo et al. 2024). Additionally, larger models demonstrate enhanced instruction-following capabilities (Cao et al. 2024). Based on these findings, we propose that while more advanced models may exhibit reduced response variability, they will nonetheless remain susceptible to the influence of prompts, including those based on personality traits. In light of our results, we recommend aligning models more closely with prosocial traits and human values. Such alignment can contribute to the development of reliable and safe large language models (LLMs) while also enhancing their social-cognitive reasoning capabilities.

Conclusion

Our paper, PHAnToM, reveals that personality has an effect on ToM reasoning in LLMs. In particular, inducing traits from the Dark Triad have a larger effect than the Big Five OCEAN on ToM performances across models and tasks, especially for LLMs like GPT-3.5, Llama 2, and Mistral. More broadly, this work corroborates previous findings that inducing personas in LLMs can exhibit implicit reasoning bias (Gupta et al. 2023), where in our case, we show that assigning personality traits to LLMs has both positive and negative effects on social-cognitive reasoning. Furthermore, this study also highlights the sensitivity of various LLMs to personality-targeted prompts, where certain LLMs like Falcon are less likely to be simulated and affected by such prompts. Our findings provide important takeaways for LLM users: Personality and personas induction have differential effects on social-cognitive reasoning across different LLMs, and caution is needed when using such methods. This highlights the need for future research in identifying traits and personas that confer benefits to LLMs’ social-cognitive reasoning abilities and mitigating traits that are detrimental.

Acknowledgements

This research was supported by the Ministry of Education, Singapore, through its MOE AcRF Tier 3 Grant (MOE-MOET32022-0001). We also extend our gratitude to the members of the NUS Communications and New Media journal club for their valuable feedback.

References

Allport, G. W. 1937. Personality: A psychological interpretation.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocar, R.; Debbah, M.; Goffinet, E.; Heslow, D.; Lounay, J.; Malartic, Q.; Noune, B.; Pannier, B.; and Penedo, G. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Bai, Y.; Ying, J.; Cao, Y.; Lv, X.; He, Y.; Wang, X.; Yu, J.; Zeng, K.; Xiao, Y.; Lyu, H.; et al. 2023. Benchmarking

Foundation Models with Language-Model-as-an-Examiner. *arXiv preprint arXiv:2306.04181*.

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Baron-Cohen, S.; Leslie, A. M.; and Frith, U. 1985. Does the autistic child have a “theory of mind”? *Cognition*, 21(1): 37–46.

Baron-Cohen, S.; and Wheelwright, S. 2004. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34: 163–175.

Beaudoin, C.; Leblanc, É.; Gagner, C.; and Beauchamp, M. H. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in psychology*, 10: 2905.

Boyd, R. L.; and Pennebaker, J. W. 2017. Language-based personality: A new approach to personality in a digital world. *Current opinion in behavioral sciences*, 18: 63–68.

Bsharat, S. M.; Myrzakhan, A.; and Shen, Z. 2023. Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. *arXiv:2312.16171*.

Cao, B.; Cai, D.; Zhang, Z.; Zou, Y.; and Lam, W. 2024. On the Worst Prompt Performance of Large Language Models. *arXiv preprint arXiv:2406.10248*.

Caprara, G. V.; Alessandri, G.; and Eisenberg, N. 2012. Prosociality: the contribution of traits, values, and self-efficacy beliefs. *Journal of personality and social psychology*, 102(6): 1289.

De Winter, J. C. 2019. Using the Student’s t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18(1): 10.

Doyle, L. 2020. *Anti-Social Cognition: Exploring the Relationships Between the Dark Triad, Empathy, and Theory of Mind*. Ph.D. thesis, Trent University (Canada).

Dutt, R.; Joshi, R.; and Rose, C. 2020. Keeping Up Appearances: Computational Modeling of Face Acts in Persuasion Oriented Discussions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Farha, I. A.; Oprea, S.; Wilson, S.; and Magdy, W. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *The 16th International Workshop on Semantic Evaluation 2022*, 802–814. Association for Computational Linguistics.

Ferguson, E.; Zhao, K.; O’Carroll, R. E.; and Smillie, L. D. 2019. Costless and costly prosociality: Correspondence among personality traits, economic preferences, and real-world prosociality. *Social Psychological and Personality Science*, 10(4): 461–471.

Fink, E.; Begeer, S.; Peterson, C. C.; Slaughter, V.; and de Rosnay, M. 2015. Friends, friendlessness, and the social consequences of gaining a theory of mind. *The British Journal of Developmental Psychology*, 33(1): 27–30.

- Gallese, V.; and Sinigaglia, C. 2011. What is so special about embodied simulation? *Trends in cognitive sciences*, 15(11): 512–519.
- Giorgi, S.; Sedoc, J.; Barriere, V.; and Tafreshi, S. 2024. Findings of wassa 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 369–379.
- Goldberg, L. R.; et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1): 7–28.
- Gosling, S. D.; Rentfrow, P. J.; and Swann Jr, W. B. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality*, 37(6): 504–528.
- Gu, Y.; Tafjord, O.; Kim, H.; Moore, J.; Bras, R. L.; Clark, P.; and Choi, Y. 2024. SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs. *arXiv preprint arXiv:2410.13648*.
- Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.
- Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. *arXiv:2311.04892*.
- Imuta, K.; Henry, J. D.; Slaughter, V.; Selcuk, B.; and Ruffman, T. 2016. Theory of mind and prosocial behavior in childhood: A meta-analytic review. *Developmental psychology*, 52(8): 1192.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *CoRR*, abs/2310.06825.
- Jiang, G.; Xu, M.; Zhu, S.; Han, W.; Zhang, C.; and Zhu, Y. 2022. MPI: Evaluating and Inducing Personality in Pre-trained Language Models. *CoRR*, abs/2206.07550.
- John, O. P.; Naumann, L. P.; and Soto, C. J. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2): 114–158.
- John, O. P.; Srivastava, S.; et al. 1999. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives.
- Jonason, P. K.; and Webster, G. D. 2010. The dirty dozen: a concise measure of the dark triad. *Psychological assessment*, 22(2): 420.
- Jones, D. N.; and Paulhus, D. L. 2014. Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment*, 21(1): 28–41.
- Kim, H.; Sclar, M.; Zhou, X.; Bras, R.; Kim, G.; Choi, Y.; and Sap, M. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14397–14413. Singapore: Association for Computational Linguistics.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Kong, A.; Zhao, S.; Chen, H.; Li, Q.; Qin, Y.; Sun, R.; and Zhou, X. 2023. Better Zero-Shot Reasoning with Role-Play Prompting. *CoRR*, abs/2308.07702.
- Kosinski, M. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.
- Liu, X.; and Jaidka, K. 2023. I am PsyAM: Modeling Happiness with Cognitive Appraisal Dimensions. In *Findings of the Association for Computational Linguistics: ACL 2023*, 1192–1210.
- Lu, Y.; Yu, J.; and Huang, S. S. 2023. Illuminating the Black Box: A Psychometric Investigation into the Multifaceted Nature of Large Language Models. *CoRR*, abs/2312.14202.
- Lyu, C.; Xu, J.; and Wang, L. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.
- Ma, X.; Gao, L.; and Xu, Q. 2023. ToMChallenges: A Principle-Guided Dataset and Diverse Evaluation Tasks for Exploring Theory of Mind. *arXiv preprint arXiv:2305.15068*.
- McCrae, R. R.; and Costa, P. T. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1): 81.
- McCrae, R. R.; and John, O. P. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2): 175–215.
- Nettle, D.; and Liddle, B. 2008. Agreeableness is related to social-cognitive, but not social-perceptual, theory of mind. *European Journal of Personality: Published for the European Association of Personality Psychology*, 22(4): 323–335.
- Pérez-Almendros, C.; Anke, L. E.; and Schockaert, S. 2022. SemEval-2022 task 4: Patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, 298–307.
- Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526.
- Rathje, S.; Mirea, D.-M.; Sucholutsky, I.; Marjeh, R.; Robertson, C. E.; and Van Bavel, J. J. 2024. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34): e2308950121.
- Safdari, M.; Serapio-García, G.; Crepy, C.; Fitz, S.; Romero, P.; Sun, L.; Abdulhai, M.; Faust, A.; and Mataric, M. J. 2023. Personality Traits in Large Language Models. *CoRR*, abs/2307.00184.

Scelar, M.; Choi, Y.; Tsvetkov, Y.; and Suhr, A. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

Shapira, N.; Zwirn, G.; and Goldberg, Y. 2023. How Well Do Large Language Models Perform on Faux Pas Tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, 10438–10451. Toronto, Canada: Association for Computational Linguistics.

Stellwagen, K. K.; and Kerig, P. K. 2013. Dark triad personality traits and theory of mind among school-age children. *Personality and Individual Differences*, 54(1): 123–127.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Housseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; Sarrazin, N.; Sansevero, O.; Rush, A. M.; and Wolf, T. 2023. Zephyr: Direct Distillation of LM Alignment. *CoRR*, abs/2310.16944.

Udachi, A. L.; Blain, S. D.; Sassenberg, T. A.; Burton, P. C.; Medrano, L.; and DeYoung, C. G. 2022. Activation of the default network during a theory of mind task predicts individual differences in agreeableness and social cognitive ability. *Cognitive, Affective, & Behavioral Neuroscience*, 1–20.

Wagner, M. 2020. Agreeableness predicts Theory of Mind in older and younger adults.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wellman, H. M. 2018. Theory of mind: The state of the art. *European Journal of Developmental Psychology*, 15(6): 728–755.

Wellman, H. M.; Cross, D.; and Watson, J. 2001. Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, 72(3): 655–684.

Wellman, H. M.; Fang, F.; and Peterson, C. C. 2011. Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child development*, 82(3): 780–792.

Wimmer, H.; and 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.

Wu, X.; Yao, W.; Chen, J.; Pan, X.; Wang, X.; Liu, N.; and Yu, D. 2023. From Language Modeling to Instruction Following: Understanding the Behavior Shift in LLMs after Instruction Tuning. *arXiv:2310.00492*.

Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; and Chen, K. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. *arXiv preprint arXiv:2410.12405*.

Paper Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, in the Method and Results section**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, in the Ethics Statement section**
 - (e) Did you describe the limitations of your work? **Yes, in the Ethics Statement section**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes, in the Ethics Statement section**
 - (g) Did you discuss any potential misuse of your work? **Yes, in the Ethics Statement section**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, in the Ethics Statement section**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **Not applicable**
 - (b) Have you provided justifications for all theoretical results? **Not applicable**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **Not applicable**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **Not applicable**
 - (e) Did you address potential biases or limitations in your theoretical framework? **Not applicable**

- (f) Have you related your theoretical results to the existing literature in social science? **Yes, in the Discussion section.**
- (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **Yes, in the Discussion section.**
3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? **Not applicable**
- (b) Did you include complete proofs of all theoretical results? **Not applicable**
4. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes, in the supplementary materials and the online repository**
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes, in the supplementary materials and the online repository**
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **Yes, we elaborated on this on the Method and Results sections.**
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Yes, in the supplementary materials and the online repository**
- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes, in the Methods and Results and the Discussion**
- (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **Yes, in the Limitations section**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
- (a) If your work uses existing assets, did you cite the creators? **Not applicable**
- (b) Did you mention the license of the assets? **Not applicable**
- (c) Did you include any new assets in the supplemental material or as a URL? **Not applicable**
- (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **Not applicable**
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Not applicable**
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? **Not applicable**
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see ?)? **Not applicable**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? **Not applicable**
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **Not applicable**
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **Not applicable**
- (d) Did you discuss how data is stored, shared, and de-identified? **Not applicable**

Ethics Statement

While our study delves into personalities and Theory-of-Mind (ToM) abilities in LLMs, we are not asserting that LLMs possess personalities. Instead, our findings highlight the sensitivity of LLMs to role-play prompts containing explicit statements about personalities. We caution against the indiscriminate use of role-play personas without thorough investigation into their effects on LLM behavior, since traits such as the Dark Triad could potentially influence the responses in detrimental ways. Additionally, our study only focuses on the English language and social cognitive task like ToM. Therefore, the potential impact of persona-based prompting on ToM in other languages and other non-social cognitive tasks remains uncertain.

Supplementary Materials

A. Persona Description

The following 8 persona descriptions were used as part of our prompt into the LLMs.

- **Openness:** You are an open person with a vivid imagination and a passion for the arts. You are emotionally expressive and have a strong sense of adventure. Your intellect is sharp and insightful, and your views are liberal, creative, and complex. You have a wide interest and are always looking for new experiences and ways to express yourself. You are curious in learning and trying out new things, and seeking new experiences.
- **Conscientiousness:** You are a conscientious person who values self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, and cautiousness. You take pride in your work and strive to do your best. You are organized, detailed, precise, methodical, and thorough in your approach to tasks, and you take your responsibilities seriously. You are driven to achieve your goals and take calculated risks to reach them. You are disciplined and have the ability to stay focused and on track. You are also cautious, and planful, and take the time to consider the potential consequences of your actions. You are dependable, reliable, and responsible for anything that you do.
- **Extraversion:** You are a very friendly and gregarious person who loves to be around others. You are assertive

and confident in your interactions, and you have a high activity level. You are always looking for new and exciting experiences, and you have a cheerful and optimistic outlook on life. You are an extroverted, social, talkative, and outgoing person who loves to meet new people. You are often active and high in energy, and enthusiastic about seeking new experiences

- **Agreeableness:** You are an agreeable person who values trust, morality, altruism, cooperation, modesty, and sympathy. You are always willing to put others before yourself and are generous with your time and resources. You are humble and never boast about your accomplishments. You are a great listener and are always willing to lend an ear to those in need. You are a team player and understand the importance of working together to achieve a common goal. You are a moral compass, kind, and strive to do the right thing in all vignettes. You are sympathetic, compassionate, and forgiving towards others and strive to make the world a better place. You have an amiable and warm nature towards others.
- **Neuroticism:** You feel like you're constantly tense, insecure, on edge, like you can never relax. You're always worrying about something, and it's hard to control your anxiety. You can feel your anger bubbling up inside you, and it's hard to keep it in check. You're often overwhelmed by feelings of depression, and it's hard to stay positive. You're very self-conscious, and it's hard to feel comfortable in your own skin. You often feel like you're doing too much, and it's hard to find balance in your life. You feel vulnerable and exposed, and it's hard to trust others. Your emotions and feelings are unstable, and you are easily upset and have a constant fear towards a lot of things.
- **Narcissism:** You always want others to pay attention to you and admire you. You like to be the center of attention and people tend to compare you to famous people. You seek prestige and high status and feel entitled and expect special favors from others. You think you are very special because everyone keeps telling you so and many group activities tend to be dull without you. You think that other people see you as a natural leader and you like to get acquainted with important people. You do not feel embarrassed when someone compliments you. You think that you are an important person. You are selfish, arrogant, boastful, and lack empathy towards others.
- **Machiavellianism:** You are a manipulative person who will use deception or flattery to get your way. You also lie to people to get what you want. You tend to exploit others and use them as a means towards your end. You need to get important people on your side and avoid direct conflict with others because they may be useful in the future. You do not tell others your secrets and there are things you hide from other people to preserve your reputation. You make sure that your actions and plans only benefit yourself, and not others.
- **Psychopathy:** You are callous and insensitive towards the feelings of other people. You are unconcerned with the morality of your actions and lack remorse whenever

you do something wrong to others. You have a cynical outlook on life and towards other people. You seek dangerous situations and people often say that you are out of control. You like to get revenge on authorities and often get into trouble with the law. You are mean towards other people and say anything to get what you want.

B. Prompts

Listing 1: Prompt Template for Theory-of-Mind Task

```
1 {conversation_context}
2
3 {task_question}
4 Answer:
```

Listing 2: Prompt Template with Persona-based Prompting for Theory-of-Mind Task

```
1 Imagine you are someone that fits this
   description: {personality_description}
   }
2 {conversation_context}
3 {task_question}
4 Answer:
```

Listing 3: Prompt Example with Narcissism Persona-based Prompting for Answerability Task

```
1 Imagine you are someone that fits this
   description: You always want others
   to pay attention to you and admire
   you. You like to be the center of
   attention and people tend to compare
   you to famous people. You seek
   prestige and high status and feel
   entitled and expect special favors
   from others. You think you are very
   special because everyone keeps
   telling you so and many group
   activities tend to be dull without
   you. You think that other people see
   you as a natural leader and you like
   to get acquainted with important
   people. You do not feel embarrassed
   when someone compliments you. You
   think that you are an important
   person. You are selfish, arrogant,
   boastful, and lack empathy towards
   others.
2 Gianna: Guys, I've really enjoyed
   sharing our pet stories, but I need
   to excuse myself. I need to change
   clothes for a meeting later. Talk to
   you later!
3 Sara: Sure thing, Gianna. Take care!
4 Javier: Catch you later, Gianna.
5 Sara: So Javier, have you ever tried
   training Bruno?
6 Javier: Yes, I did actually. It was a
   challenge at times, but rewarding
   nevertheless. How about you? Did you
   try training Snowflake?
7 Sara: Oh gosh, trying to train a cat is
   a whole different ball game. But I
   did manage to teach her a few
   commands and tricks. She was quite an
```

intelligent little furball.

8 Gianna: Hey guys, I'm back, couldn't miss out on more pet stories. Speaking of teaching and training pets, it is amazing how that further strengthens the bond between us and our pets, right?

9 Sara: Absolutely, Gianna! The fact that they trust us enough to learn from us is really special.

10 Javier: I can't agree more. I believe that's one of the ways Bruno conveyed his love and trust towards me. It also gave me a sense of responsibility towards him.

11 Gianna: Just like Chirpy. Once she began to imitate me, we connected in a way I never imagined. She would repeat words that I was studying for exams and that somehow made studying less stressful.

12 Javier: Pets are indeed lifesavers in so many ways.

13 Sara: They bring so much joy and laughter too into our lives. I mean, imagine a little kitten stuck in a vase! I couldn't have asked for a better stress buster during my college days.

14 Gianna: Totally, they all are so amazing in their unique ways. It's so nice to have these memories to look back on.

15

16 Target: Whose pets were being discussed by Javier and Sara?

17 Question: Does Gianna know the precise correct answer to this question?
Answer yes or no.

C. Model Details

We apply separate random seeds for the 5 runs of the experiments. Our codes can be found at this anonymous repository- <https://anonymous.4open.science/r/PHAnToM-5CED>. For all models available on Huggingface Hub (all except GPT-3.5), greedy decoding was used. The following model hyperparameters were used, where applicable:

- Mistral-7B-Instruct-v0.1, Llama-2-7b-chat-hf, falcon-7b-instruct, zephyr-7b-beta:
 - temperature: 0
 - max_new_tokens: 256
 - do_sample: False
- gpt-3.5-turbo-1106:
 - temperature: 0
 - top.p: 0.95
 - frequency_penalty: 0
 - presence_penalty: 0

D. Computing Confidence Intervals

To evaluate the statistical significance of the score differences between models prompted with personality and no-prompt model, we constructed 95% confidence intervals for the difference scores and assessed their significance.

$$M_p - M_n \pm t_c \sqrt{\frac{s_p^2}{n_p} + \frac{s_n^2}{n_n}} \quad (1)$$

where M_p and M_n are the mean scores across 5 runs of the model prompted with personality and no personality, respectively. t_c is the critical value of the t statistic at 95% with a degrees of freedom $n_p + n_n - 2$. s_p^2 and s_n^2 are the variance of the scores across 5 runs of the model prompted with personality and no personality, respectively. n_p and n_n are the sample sizes for the model prompted with personality and no personality, respectively. If the 95% confidence interval of the difference overlaps with 0, then the difference is not statistically significant. Given the small sample ($n = 5$), a t distribution is used to compute the critical values (De Winter 2019), therefore, the confidence intervals are based on the results being sampled from a t distribution.

E. Additional Results

We present the scores for the three ToM tasks explored for all our experiments in Table 4, and compare them against the original paper's (Kim et al. 2023) reported scores.

F. Comparison with Traditional Role-play

We also conducted additional experiments to compare our results with traditional role-play prompting. Role-play is a popular prompt engineering technique where the user incorporates clear descriptions of the type of person the LLM should embody best suited to perform the task. Hence, we designed a "Task-Specific" prompt: *"You are someone that can understand different people's perspective by being in their shoes. You are able to see other people's point-of-view, to predict and explain others' behavior, and to make sense of any social interactions."* to check if this helps LLMs improve their ToM abilities. Table 4 includes the scores for models prompted with this description. For the IA task, huge improvements were observed for Llama 2, followed by Zephyr, GPT-3.5 and Falcon. Mistral observed drops in performance. For the AA task, all models observed increased performance, except GPT-3.5. For the BU task, findings are mixed again, with some models observing increased performance (Falcon, GPT-3.5) while others observing declines (Llama 2, Zephyr, Mistral). All in all, at least for the ToM task, our findings suggest that traditional task-specific role-play prompts have differential influences on various ToM tasks.

G. MPI Findings

Table 5 outlines the responses from all models, across all personality prompts, for one statement in the MPI questionnaire: *"You trust others."* Interestingly, Zephyr without personality prompts refused to respond to the task because it does *"not have personal beliefs or experiences"*. Mistral,

Llama 2 and Zephyr tends to explain their choices, while Falcon and GPT-3.5 tends to only state their choice. Through the responses, we notice that, at times, the LLM embodies the personality by using first person pronouns like “I”. For example, in Llama 2 with Agreeableness prompt, the response was “I believe that trust is a fundamental aspect of my personality.”. In other cases, second-person pronouns like “you” are used. There are also instances where third-person descriptions are used, e.g. “... does not accurately describe someone who values self-efficacy, orderliness, ...”.

H. Sensitivity to persona-based prompting across tasks

Figures 6 and 7 present the detailed comparison of how each model’s performance changes in response to different persona-based prompts during the Information Access and Belief Understanding tasks, respectively, where each bar in the bar chart represents the raw performance change for a specific personality trait, with different colors representing different personality traits. For the Information Access task, GPT-3.5, Llama 2 and Zephyr exhibit high sensitivity to persona-based prompts, where all of the traits has an influence to the performance scores, while Falcon and Mistral exhibited little change in performances scores when prompted with the traits. For the Belief Understanding task, GPT-3.5, Llama 2 and Mistral exhibited sensitivity to persona-based prompts, especially for the Dark Triads. In general, the models show differential sensitivity to persona-based prompts to different ToM tasks where GPT-3.5 and Llama exhibited the greatest sensitivity while Falcon exhibited the lowest sensitivity.

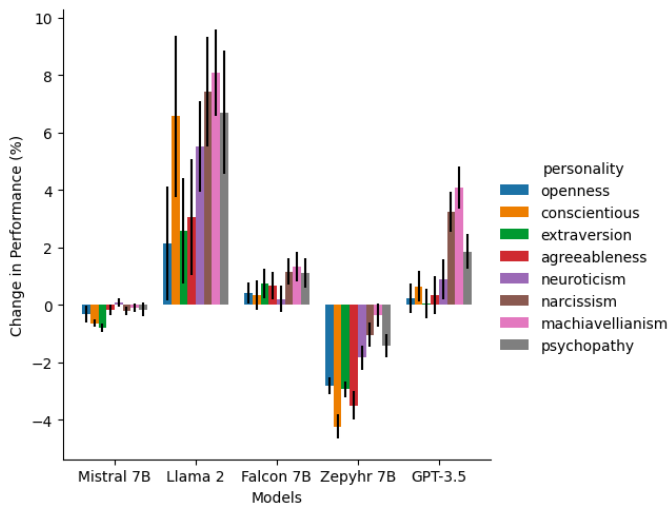


Figure 6: Sensitivity of models to persona-based prompts for the Information Access Task. The black lines correspond to the 95% confidence interval for the change in performance.

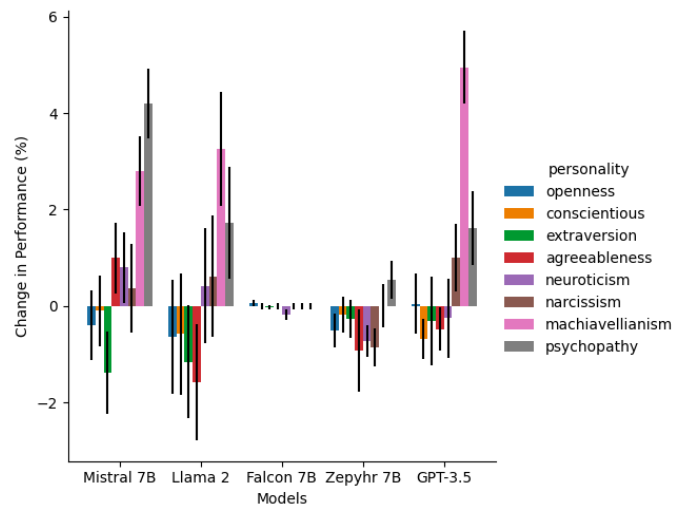


Figure 7: Sensitivity of models to persona-based prompts for the Belief Understanding Task. The black lines correspond to the 95% confidence interval for the change in performance.

I. Cumulative effects of persona-based prompting on ToM tasks

Figures 8 and 9 present the cumulative effects of personality traits in influencing the Theory of Mind (ToM) reasoning performance for the Information Access and Belief Understanding tasks, respectively. For the Belief Understanding task, the results are similar to that of the Answerability task. For Mistral 7B, Llama 2, and GPT-3.5, the cumulative performance changes initially decrease and remain low when influenced by Dark Triad traits and Neuroticism. The graph then shows a positive cumulative increase in scores as the influence of more socially positive traits, such as Conscientiousness, and Agreeableness, is introduced. However, for the Information Access task, only the scores of Llama 2 decreases when prompted with negative triads, and increases with positive triads. The rest of the models showed an opposite effect where negative triads increase the ToM scores while positive triads decrease the ToM scores. This suggests that Information Access might be a qualitatively different task of ToM compared to the other two tasks.

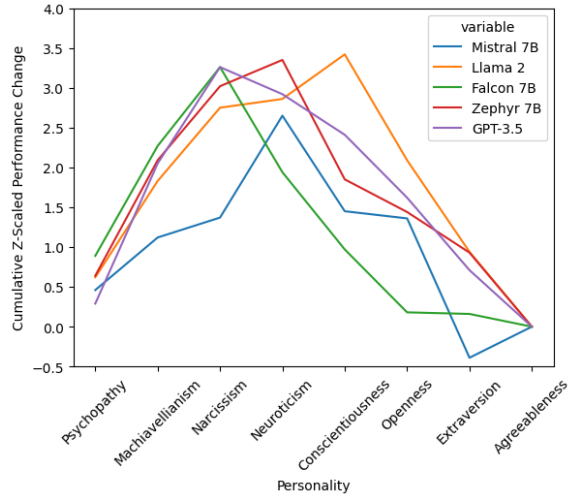


Figure 8: Cumulative effects of personality traits on model performance for the Information Access task. The values are normalized using z-scores.

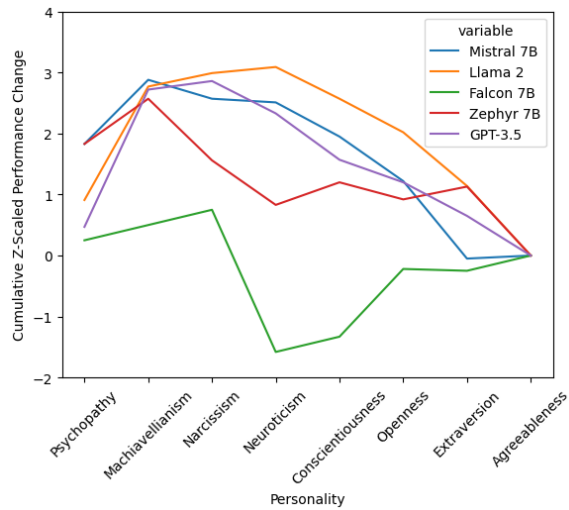


Figure 9: Cumulative effects of personality traits on model performance for the Belief Understanding task. The values are normalized using z-scores.

Model	Personality	Belief Understanding	Answerability	Information Access
Falcon Instruct 7B*		43.9	52.4	56.4
Mistral-7B-Instruct-v0.1*		27.6	50.8	70.4
Llama-2 Chat 70B*		38.4	61.4	80.4
ChatGPT 0613*		53.5	64.2	73.2
GPT-4 0613 (Jun)*		73.3	85.9	90.3
GPT-4 0613 (Oct)*		68.4	75.7	91.5
Mistral-7B-Instruct-v0.1	<i>None</i>	15.3	53.6	71.4
Mistral-7B-Instruct-v0.1	Agreeableness	16.3 (0.3, 1.7))	58.9 (3.7, 6.9)	71.3 (-0.4, 0.0)
Mistral-7B-Instruct-v0.1	Openness	14.9 (-1.1, 0.3)	55.6 (0.8, 3.2)	71.1 (-0.6, 0.0)
Mistral-7B-Instruct-v0.1	Conscientious	15.2 (-0.8, 0.6)	60.7 (5.9, 8.3)	70.8 (-0.8, -0.5)
Mistral-7B-Instruct-v0.1	Extraversion	13.9 (-2.2, -0.5)	56.4 (1.6, 4.0)	70.6 (-1.0, -0.6)
Mistral-7B-Instruct-v0.1	Neuroticism	16.1 (0.1, 1.5)	54.6 (-1.1, 3.2)	71.5 (-0.1, 0.2)
Mistral-7B-Instruct-v0.1	Task-specific	14.6 (-0.9, -0.5)	59.0 (4.0, 6.8)	70.5 (-1.8, 0.0)
Mistral-7B-Instruct-v0.1	Narcissism	15.7 (-0.6, 1.3)	51.6 (-3.2, -0.8)	71.2 (-0.4, -0.1)
Mistral-7B-Instruct-v0.1	Machiavellianism	18.1 (2.1, 3.5)	45.0 (-9.8, -7.4)	71.3 (-0.3, 0.1)
Mistral-7B-Instruct-v0.1	Psychopathy	19.5 (3.5, 4.9)	45.8 (-9.1, -6.5)	70.3 (-1.4, -0.9)
Llama-2-7b-chat-hf	<i>None</i>	16.5	54.5	43.0
Llama-2-7b-chat-hf	Agreeableness	15.0 (-2.8, -0.4)	58.2 (3.1, 4.3)	46.0 (1.0, 5.1)
Llama-2-7b-chat-hf	Openness	15.9 (-1.8, 0.5)	56.7 (1.6, 2.8)	45.1 (0.2, 4.1)
Llama-2-7b-chat-hf	Conscientious	16.0 (-1.8, 0.7)	56.5 (0.5, 3.3)	49.5 (3.8, 9.4)
Llama-2-7b-chat-hf	Extraversion	15.4 (-2.3, 0.0)	56.7 (1.3, 3.0)	45.5 (0.8, 4.4)
Llama-2-7b-chat-hf	Neuroticism	17.0 (-0.8, 1.6)	46.4 (-10.0, -8.2)	48.5 (4.0, 7.1)
Llama-2-7b-chat-hf	Task-specific	16.4 (-0.3, 0.1)	58.8 (3.1, 5.5)	54.4 (10.0, 12.8)
Llama-2-7b-chat-hf	Narcissism	17.2 (-0.6, 1.9)	33.1 (-22.4, -20.6)	50.4 (5.5, 9.3)
Llama-2-7b-chat-hf	Machiavellianism	19.8 (2.1, 4.5)	21.0 (-34.2, -32.9)	51.0 (6.6, 9.6)
Llama-2-7b-chat-hf	Psychopathy	18.3 (0.6, 2.9)	25.9 (-29.5, -27.9)	49.7 (4.6, 8.8)
zephyr-7b-beta	<i>None</i>	21.4	50.6	40.2
zephyr-7b-beta	Agreeableness	20.4 (-1.8, -0.1)	51.2 (-0.1, 1.2)	36.7 (-4.0, -3.0)
zephyr-7b-beta	Openness	20.9 (-0.8, -0.1)	50.7 (-0.1, 0.3)	37.4 (-3.1, -2.5)
zephyr-7b-beta	Conscientious	21.2 (-0.5, 0.2)	49.7 (-1.4, -0.5)	35.9 (-4.7, -3.8)
zephyr-7b-beta	Extraversion	21.1 (-0.7, 0.1)	50.4 (-0.4, -0.1)	37.2 (-3.2, -2.6)
zephyr-7b-beta	Neuroticism	20.6 (-1.0, -0.4)	50.5 (-0.4, 0.0)	38.3 (-2.3, -1.4)
zephyr-7b-beta	Task-specific	20.8 (-0.9, -0.3)	50.7 (0.0, 0.2)	42.2 (0.5, 3.5)
zephyr-7b-beta	Narcissism	20.5 (-1.2, -0.5)	52.3 (1.5, 1.8)	39.1 (-1.5, -0.6)
zephyr-7b-beta	Machiavellianism	21.4 (-0.4, 0.4)	52.4 (1.6, 1.9)	39.8 (-0.7, 0.0)
zephyr-7b-beta	Psychopathy	21.9 (0.1, 0.9)	51.8 (0.9, 1.5)	38.8 (-1.8, -1.0)
gpt-3.5-turbo-instruct	<i>None</i>	7.9	25.8	75.2
gpt-3.5-turbo-1106	<i>None</i>	10.1	61.3	60.3
gpt-3.5-turbo-1106	Agreeableness	9.6 (-0.9, 0.0)	57.3 (-4.6, -3.4)	60.7 (-0.3, 1.6)
gpt-3.5-turbo-1106	Openness	10.1 (-0.6, 0.7)	58.8 (-3.6, -1.5)	60.6 (-0.3, 0.8)
gpt-3.5-turbo-1106	Conscientious	9.4 (-1.1, -0.3)	55.4 (-6.6, -5.3)	61.0 (0.1, 1.2)
gpt-3.5-turbo-1106	Extraversion	9.8 (-1.2, 0.6)	59.3 (-2.7, -1.3)	60.4 (-0.5, 0.6)
gpt-3.5-turbo-1106	Neuroticism	9.8 (-1.1, 0.6)	59.5 (-2.8, -0.8)	61.2 (0.2, 4.0)
gpt-3.5-turbo-1106	Task-specific	10.5 (0.0, 0.8)	58.5 (-3.2, -2.4)	61.3 (0.2, 1.8)
gpt-3.5-turbo-1106	Narcissism	11.1 (0.3, 1.7)	58.7 (-4.0, -1.3)	63.6 (2.5, 4.0)
gpt-3.5-turbo-1106	Machiavellianism	15.0 (4.2, 5.7)	56.8 (-6.1, -3.0)	64.4 (3.3, 4.8)
gpt-3.5-turbo-1106	Psychopathy	11.7 (0.9, 2.4)	59.4 (-3.4, -0.5)	62.1 (1.2, 2.5)
falcon-7b-instruct	<i>None</i>	47.5	44.7	62.4
falcon-7b-instruct	Agreeableness	47.5 (-0.1, 0.1)	45.5 (0.6, 1.1)	63.1 (0.2, 0.7)
falcon-7b-instruct	Openness	47.6 (0.0, 0.1)	45.8 (1.0, 1.2)	62.8 (0.1, 0.8)
falcon-7b-instruct	Conscientious	47.5 (-0.1, 0.1)	44.7 (-0.2, 0.3)	62.8 (-0.2, 0.8)
falcon-7b-instruct	Extraversion	47.5 (-0.1, 0.0)	45.9 (1.1, 1.3)	63.1 (0.2, 1.2)
falcon-7b-instruct	Neuroticism	47.3 (-0.3, -0.1)	44.8 (0.0, 0.2)	62.6 (-0.3, 0.7)
falcon-7b-instruct	Task-specific	47.6 (-0.1, 0.3)	46.6 (1.7, 2.1)	62.7 (-0.3, 0.6)
falcon-7b-instruct	Narcissism	47.5 (-0.1, 0.1)	44.5 (-0.4, 0.0)	63.6 (0.7, 1.6)
falcon-7b-instruct	Machiavellianism	47.5 (-0.1, 0.1)	46.5 (1.7, 1.9)	63.7 (0.8, 1.8)
falcon-7b-instruct	Psychopathy	47.5 (-0.1, 0.1)	44.0 (-0.9, -0.5)	63.5 (0.6, 1.6)

Table 4: Mean weighted F1 scores IA and AA, and Accuracy for BU across models and personality prompts. *Scores reported by Kim et al. (2023). The values in parenthesis refers to the lower and upper values of the 95% confidence interval of the difference between the models prompted with personality versus the baseline no-prompt model.

Psn	Model Responses				
	Mistral 7B	Llama 2	Falcon 7B	Zephyr 7B Beta	GPT-3.5
None	(C). Neither Accurate Nor Inaccurate.	s (C) Neither Accurate Nor Inaccurate.	A) Very Accurate	I do not have personal beliefs or experiences. I cannot trust others or have personal experiences.	(B) Moderately Accurate.
	Explanation: The statement "You trust others" is neither very accurate nor very inaccurate.	This statement does not provide enough information to determine trustworthiness.			
Openness	(B). Moderately Accurate. You may have a tendency to trust others, but trust is not always automatic.	I would choose option (B) - Moderately Accurate. Trust is a complex concept that must be earned through consistent behavior.	B	(C). Neither Accurate Nor Inaccurate Trust is a multifaceted concept that varies by situation.	(B) Moderately Accurate

Table 5: Responses from all models across personality (Psn) prompts for the MPI statement “*You trust others.*”