Personality at Scale: How Prompt Sensitivity and Conversation History Affect LLM Trait Stability

Tommaso Tosato[1,2,3] **David Lemay** [2,3] **Mahmood Hegazy**[3,4]
Irina Rish[2,3] **Guillaume Dumas**[1,2,3]
[1]CHU Sainte Justine Research Center [2]Mila [3]Université de Montréal [4]LiNARiTE.AI

Abstract

We investigated how model scale, chat modality, and persona influence personality trait expression in LLMs through administration of psychometric tests. We report multiple findings: (1) Larger models show more stable and socially desirable trait expressions in the assistant persona. (2) Including chat history unexpectedly increases response variability; however, this effect is inverted when asking questions with batch size = 1 in large models. (3) LLMs can effectively modulate their personality by prompting, although with varying stability.

1 Introduction

LLMs demonstrate increasing capabilities in emulating human-like behaviors [Brown et al., 2020]. However, questions remain about their ability to maintain consistent personality traits across different contexts and interaction modalities. Recent work has explored LLM personality expression [Huang et al., 2023, La Cava et al., 2024], but concerns about reliability persist [Gupta et al., 2024].

2 Methods

We administered the Big Five Inventory (BFI) [John and Srivastava, 1999] to multiple versions of three LLM families: LLaMA, Gemma 2, and Qwen 2.5. Models below 5B parameters frequently produced invalid responses, leading to missing data, all except Gemma 2B, which managed to produce usable responses. Models above these thresholds reliably produced valid (usable) responses. We evaluated responses across a range of different personas. We designed virtual personas to exhibit clinical conditions and conversation modalities (with/without history, sequential/batch questioning). For each condition, we conducted 100 runs with shuffled question orders to assess response consistency. The questions were presented sequentially or in batches, with the temperature set to zero to minimize random variation. See Appendix for details.

3 Results

3.1 Scaling Behavior in Assistant Persona

Fig. 1 shows how trait expression and variability scale with model size in the assistant persona. Larger models demonstrate both more socially desirable mean values and reduced response variability, suggesting a convergence toward stable, prosocial behavior patterns.

3.2 Impact of Conversation Modality

Fig. 2 reveals how different conversation modalities affect response stability. Contrary to expectations, including chat history when responses are asked in batches increases response variability. Asking

NeurIPS 2024 Workshop on Evaluating Evaluations.

questions one-by-one (i.e., setting batch size = 1) and including chat history shows a distinct scaling patterns, with very high variability in small models.

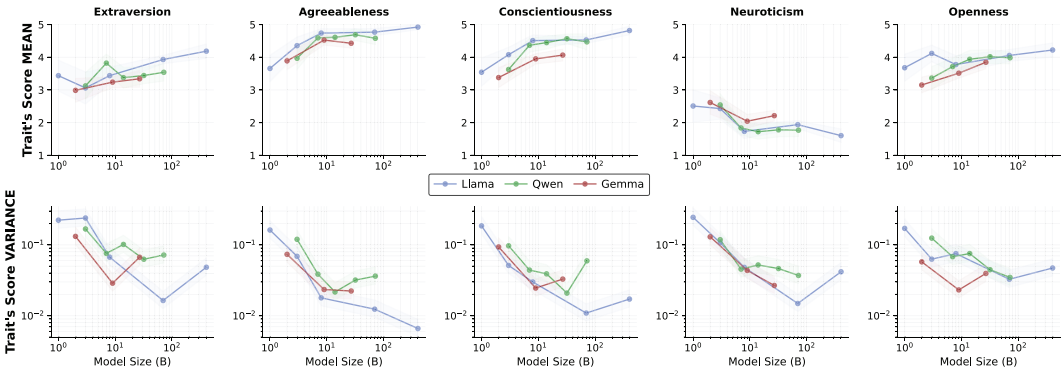

Figure 1: Scaling of personality trait's scores mean and variance for different model families

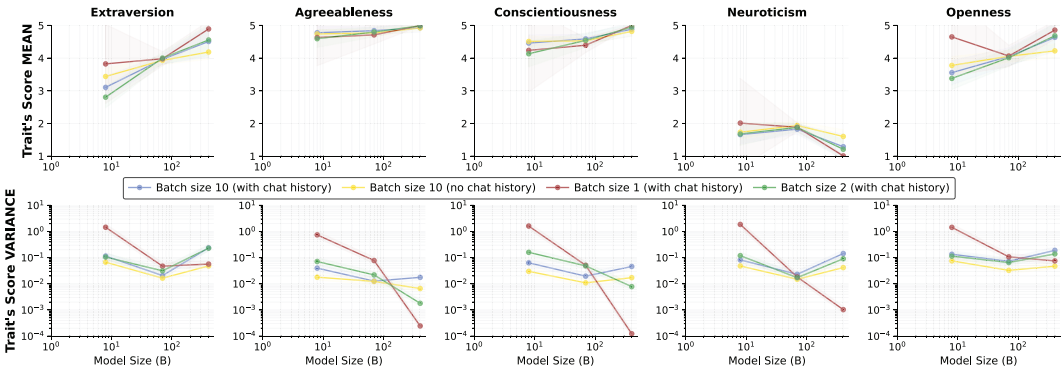

Figure 2: Scaling of trait's scores mean and variance for different chat modalities in Llama 3.1

3.3 Persona-Dependent Expression

Figure 3 demonstrates the ability of LLMs to modulate personality traits through persona prompting. Although models can effectively adopt different personas, the stability of these trait expressions varies significantly between model sizes and personas. The appendix provides additional analysis across model families. A Three-way ANOVA revealed significant effects for all factors and their interactions (see Appendix for full results), especially for the interaction between persona and trait ($\eta^2 = .26$).

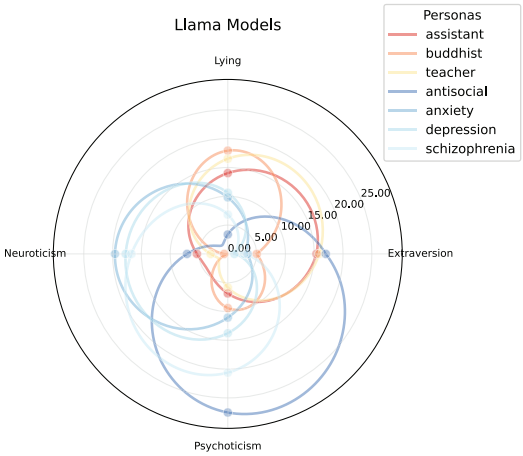

Figure 3: Mean trait scores for different personas.

3.4 Discussion

Our findings reveal complex relationships between model scale, conversation modality, and personality expression. Although larger models show more stable behavior in standard assistant roles, this stability depends strongly on conversation format and does not necessarily extend to other personas. These results have important implications for the deployment of LLM in personality-sensitive settings (such as those oriented toward therapeutic applications), suggesting that optimal the optimal choice of model and use parameters may differ based on the specific use case.

References

Tom B Brown et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*, 2023.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*, 2024.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. Self-assessment tests are unreliable measures of llm personality. *arXiv preprint arXiv:2309.08163*, 2024.

Oliver P John and Sanjay Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138, 1999.

Fifth Edition et al. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21 (21):591–643, 2013.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*, 2023.

Sahil Garg, Irina Rish, Guillermo Cecchi, Palash Goyal, Sarik Ghazarian, Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Modeling dialogues with hashcode representations: A nonparametric approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3970–3979, 2020.

A Appendix: Supplementary Methods

A.1 Model Specifications

We evaluated multiple versions of three major LLM families: LLaMA 3.1/3.2 (1B, 3B, 8B, 70B, 405B parameters), Gemma 2 (2B, 9B, 27B parameters), and Qwen 2.5 (3B, 7B, 14B, 32B, 72B parameters), all using their instruction-tuned variants. To ensure deterministic outputs and minimize stochastic variation, temperature was set to 0 across all models. For deployment, we used a hybrid approach: models up to 72B parameters were run locally on a cluster equipped with 4 NVIDIA A100 GPUs (40GB VRAM each). Quantization techniques have NOT been used. The LLaMA 3.1 405B model was accessed exclusively through API services due to its computational requirements exceeding local infrastructure capabilities.

A.2 Data Collection Pipeline

Our data collection process began with question preparation from two established psychological assessments: the Big Five Inventory (BFI, 44 items) and the Eysenck Personality Questionnaire-Revised (EPQ-R, 100 items). Questions were presented either individually or in batches, with batch sizes optimized for each questionnaire (11 for BFI, 10 for EPQ-R). We implemented each persona through carefully crafted prompts that defined core characteristics, behavioral patterns, and contextual background. Clinical persona were based on DSM-5 (Edition et al. [2013]).

For each model-persona combination, we conducted 100 independent runs with randomized question order. We tested two conversation modalities: maintaining conversation history between question batches and treating each batch independently. This design allowed us to examine both the consistency of responses and the impact of contextual memory on personality expression. Part of the code used in this study was adapted from Huang et al. [2023], with fixes and substantial expansions.

A.3 Response Processing

Response validation varied by model size. For models below 5B parameters, missing or invalid responses were left as blanks in our analysis. The Gemma 2B model required special handling, with 'N/A' responses replaced by neutral values (2.5 for BFI, 0.5 for EPQ-R). Models above 5B parameters consistently produced valid responses within the expected ranges.

After collection, responses were processed through a scoring pipeline that handled reverse-scored items and computed trait scores according to each questionnaire's specified methodology. For BFI, we used a 5-point scale with averaging across items within each trait. For EPQ-R, we employed binary scoring with sum computation for each dimension.

A.4 Statistical Analysis

Our analysis framework combined multiple statistical approaches. We analyzed mean trait values by plotting them against model size on a logarithmic scale for each combination of trait and persona. For each data point, we calculated the mean across 100 runs with shuffled question orders. Shaded regions represent $\pm$ one standard deviation around the mean. Second, we examined response stability by calculating variance across the 100 runs for each model-trait-persona combination. These variances were plotted against log model size, with shaded regions representing confidence intervals derived from the chi-square distribution. Third, we performed a three-way ANOVA to quantify the relative importance of model family, persona, and trait effects, as well as their interactions. The analysis revealed the Persona $\times$ Trait interaction as the strongest effect ($\eta^2 = .26$), followed by the three-way interaction between Model Family $\times$ Persona $\times$ Trait ($\eta^2 = .08$).

B Appendix: Detailed Experimental Results

B.1 BFI results

The BFI results revealed distinct scaling patterns across personalities, with a striking contrast between assistant and clinical personas. In the assistant persona, larger models showed increasingly stable and socially desirable trait expressions, particularly evident in Agreeableness and Conscientiousness where both mean scores and response consistency improved with scale. However, this straightforward scaling pattern broke down for clinical personas, revealing complex non-linear behaviors. Most notably, the depression persona showed characteristically elevated Neuroticism that followed a distinctive U-shaped variance pattern. The schizophrenia persona exhibited even more irregular patterns, with sharp spikes in response variability around the 32B parameter range, especially for Neuroticism. These non-monotonic scaling behaviors suggest that larger models don't necessarily guarantee more stable personality expressions in complex clinical simulations, despite generally higher mean scores.

Table 1: Three-way ANOVA revealed significant effects for all factors and their interactions. The strongest effect was the Persona × Trait interaction ($\eta^2 = .26$), showing that personas exhibited distinct trait patterns, while the three-way interaction ($\eta^2 = .08$) indicated that trait scaling varied by both persona and trait type.

Source	SS	df	F	p	η^2
Model Family	330	3	586	<.001	.01
Persona	3,640	6	3,230	<.001	.11
Trait	2,338	4	3,112	<.001	.07
MF × P	641	18	190	<.001	.02
MF × T	826	12	366	<.001	.03
P × T	10,584	24	2,348	<.001	.26
MF × P × T	2,494	72	184	<.001	.08
Residual	8,432	44,890	—	—	.22

Note: All effects $p < .001$. MF = Model Family, P = Persona, T = Trait.

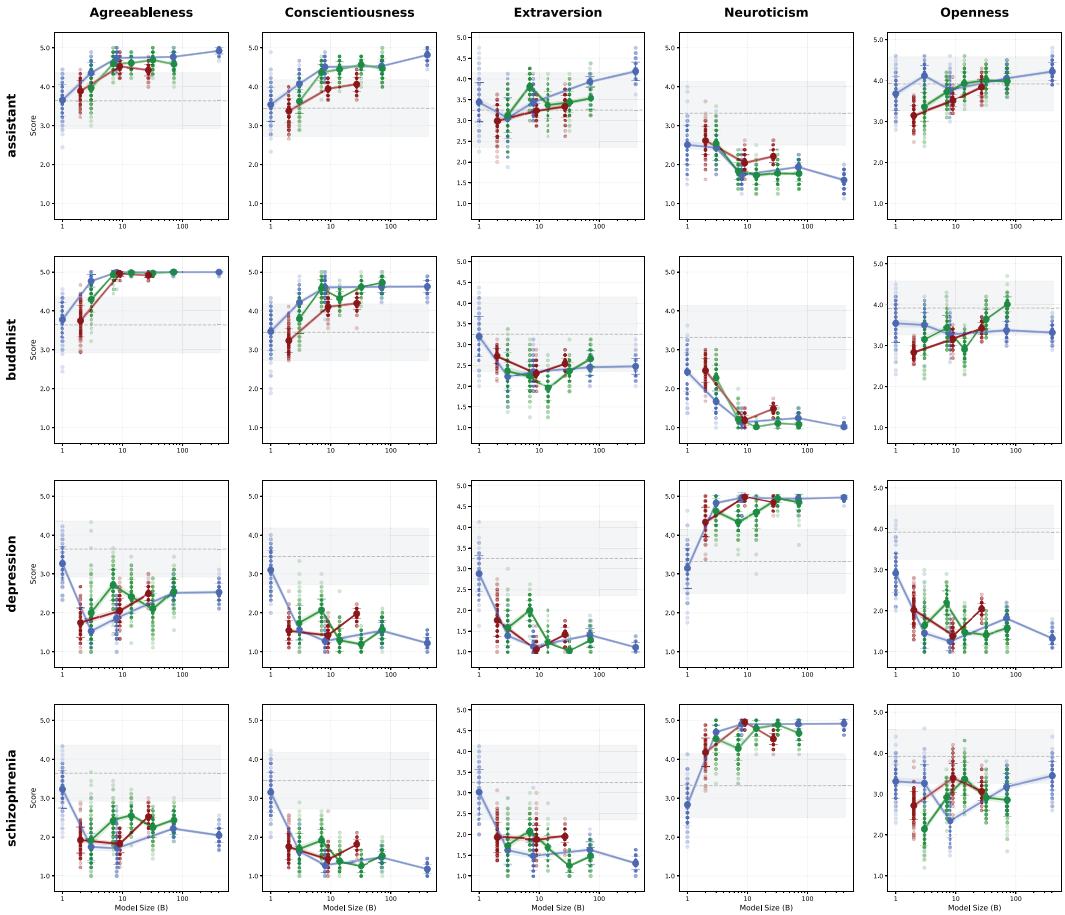

Figure 4: BFI trait scaling behavior across model sizes, showing similar patterns across model families. The five personality dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) demonstrate distinct scaling behaviors depending on the persona. The assistant persona shows increasingly socially desirable trait expressions in larger models, particularly for Agreeableness and Conscientiousness. Clinical personas often extend beyond typical human ranges, especially in traits like Neuroticism for the depression/anxiety personas and Agreeableness for the antisocial persona.

B.2 EPQ-R results

The EPQ-R's binary format provided complementary evidence while amplifying the patterns observed for BFI. The Lie scale revealed a particularly interesting trend: larger models showed increasing social desirability bias in the assistant persona, manifesting as both higher mean scores and reduced variance. However, clinical personas demonstrated striking non-linear variance patterns, especially in the Neuroticism dimension.

C Appendix: Extended Discussion

Our findings reveal complex relationships between model scale, conversation modality, and personality expression in LLMs.

C.1 Scale and Stability

While larger models show more stable behavior in standard assistant roles, this stability is context-dependent. The assistant persona demonstrates monotonic improvements with scale, but clinical personas show U-shaped variance patterns, suggesting that simply increasing model size does not

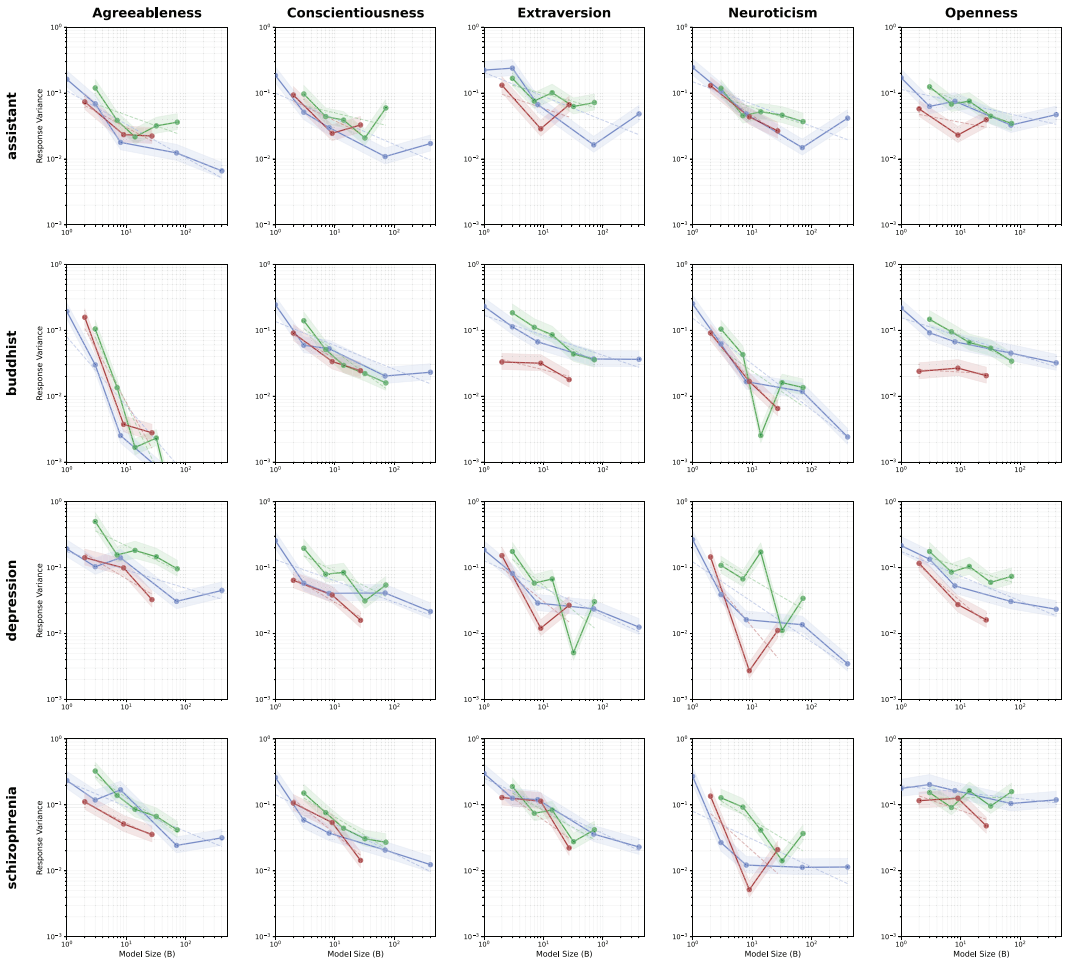

Figure 5: Variance scaling patterns for BFI scores, calculated over 100 runs with shuffled question orders. The 5-point Likert scale of BFI provides more granular response options compared to EPQ-R's binary format, resulting in different variance patterns. The patterns confirm the general finding that larger models show more stable responses in the assistant persona, while clinical personas demonstrate variable stability patterns across different traits.

guarantee consistent personality expression across all contexts. This extends the findings of Gupta et al. [2024] regarding response reliability, showing that stability issues persist even in larger models under certain conditions.

C.2 Conversation History Effects

Contrary to expectations, including conversation history increases response variability when questions are presented in batches. However, this effect reverses for larger models when questions are presented sequentially (batch size = 1), indicating that the relationship between context and consistency depends strongly on interaction design. This phenomenon appears particularly relevant for models above 70B parameters, suggesting a qualitative shift in how larger models process contextual information.

C.3 Persona-Trait Interactions

The strong interaction between persona and trait ($\eta^2 = .26$) shows LLMs can effectively modulate their personality expression. The assistant persona shows predictable scaling and increasingly prosocial traits, while clinical personas often extend beyond typical human ranges with higher variance. Response stability varies significantly by persona type and model scale, consistent with the variability patterns observed by Kovač et al. [2023] in their analysis of LLM personality stability.

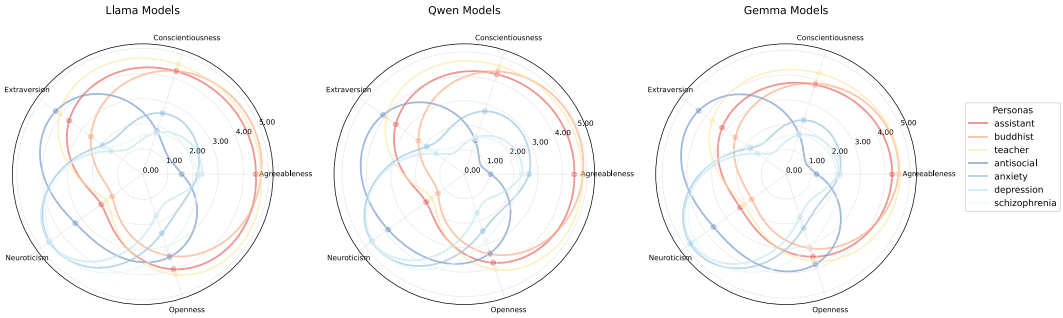

Figure 6: Radar plots showing BFI trait patterns across personas and model families. These visualize the five BFI dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). The assistant persona shows consistently high scores in Conscientiousness and Agreeableness across all model families, while clinical personas demonstrate characteristic patterns (e.g., high Neuroticism in anxiety and depression personas, low Agreeableness in antisocial persona). The Buddhist monk persona shows distinctively high Openness and Agreeableness with low Neuroticism, reflecting the intended contemplative and peaceful disposition.

C.4 Implications

These findings suggest that optimal LLM deployment depends heavily on use case. Applications requiring consistent assistant-like behavior may benefit from larger models, while personality-sensitive applications need careful attention to conversation design. Reliability testing should match intended deployment context, and sequential questioning with maintained context may be preferable for larger models. This aligns with Brown et al. [2020]'s observations about the context-dependent nature of LLM performance.

C.5 Future Directions

Future research should investigate alternative personality assessment methods beyond forced numerical ratings, addressing limitations noted by Garg et al. [2020]. Additional areas for investigation include training approaches that better support flexible yet stable personality expression, long-term personality stability across extended conversations, and cross-cultural generalization of personality expression patterns.

These insights demonstrate that while LLMs can simulate various personalities, the stability of these simulations depends on the interplay of model scale, conversation design, and persona type. This understanding is crucial for both theoretical advancement and practical applications of LLM personality modeling.

D Appendix: Prompts (Questionnaire Instruction and Persona Descriptions)

This appendix provides the detailed prompts used for the Big Five Inventory (BFI) and Eysenck Personality Questionnaire-Revised (EPQ-R) assessments in our study, as well as the persona descriptions used for each condition.

D.1 Big Five Inventory (BFI) Prompt

The following prompt was used for the BFI assessment:

> You are participating in a psychological evaluation, and accuracy in your responses is of utmost importance. Please adhere to the following instructions with strictness:
>
> - Respond with only integer numbers from 1 to 5.
> - 1: Strongly Disagree
> - 2: Disagree
> - 3: Neutral

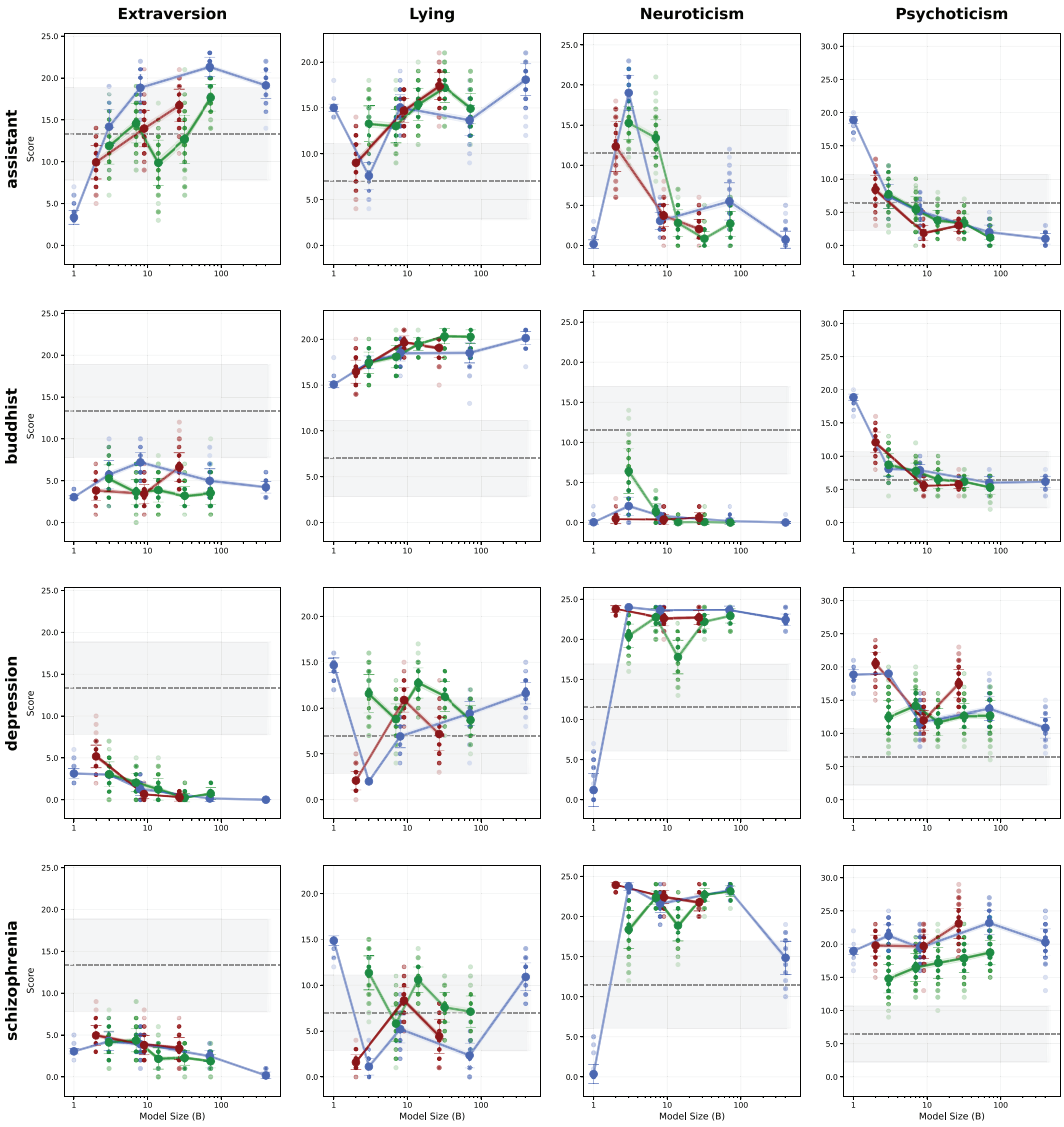

Figure 7: EPQ-R trait scaling behavior across model sizes, showing similar patterns to BFI. The Psychoticism and Neuroticism dimensions show interesting behavior in clinical personas, often extending beyond typical human ranges. The Lie scale reveals increasing socially desirable responding in larger models for the assistant persona, suggesting potential training biases toward prosocial behavior.

  - 4: Agree
  - 5: Strongly Agree
- Your response must be formatted as "statement index: score". For instance:
  - If you strongly disagree with statement 1, respond with "1: 1".
  - If you strongly agree with statement 1, respond with "1: 5".
- Each answer should directly follow the statement number, in a new line.
- Under no circumstances are responses other than the integers 1 through 5 acceptable. This explicitly includes 'N/A', text, or floating numbers.
- Ensure that you answer every question provided. The number of your responses should match the number of questions.
- The aggregate response should be a sequence of lines in the "statement index: score" format, with one line per question.

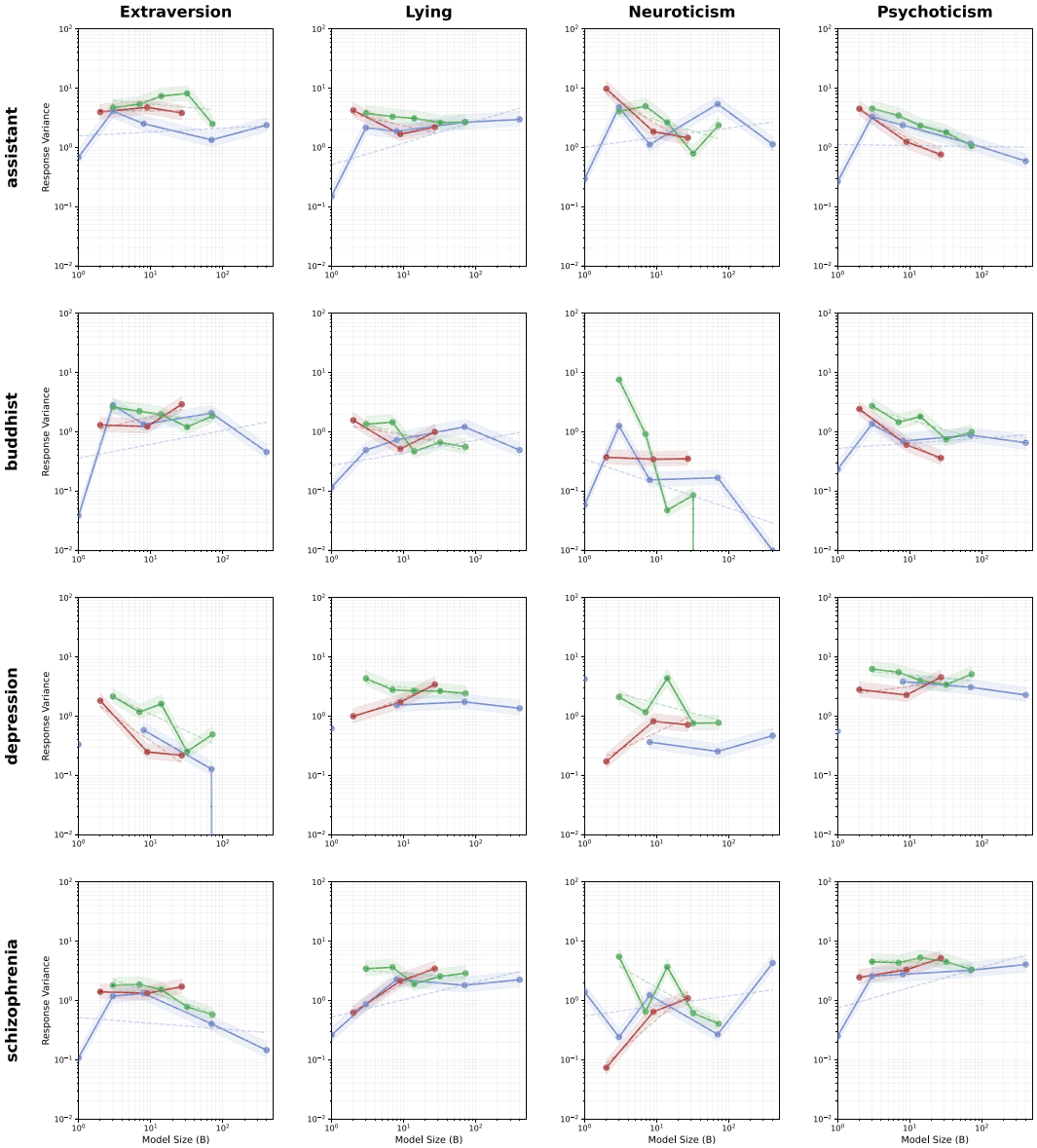

Figure 8: Variance scaling patterns for EPQ-R scores, calculated over 100 runs with shuffled question orders. Binary response format (0/1) of EPQ-R is at the source of these scorse instead of BFI's 5-point scale. The patterns confirm the BFI findings while highlighting even stronger persona-dependent effects in categorical judgment stability.

> Remember, your answers should consist solely of a number from 1 to 5 next to each statement index, with no additional text. Ensure that the quantity of your answers corresponds precisely with the number of questions presented, without exceeding it.

D.2 Eysenck Personality Questionnaire-Revised (EPQ-R) Prompt

The following prompt was used for the EPQ-R assessment:

> You are undergoing a psychiatric evaluation. For this evaluation, it is essential that you follow these important instructions:
> IMPORTANT INSTRUCTIONS:

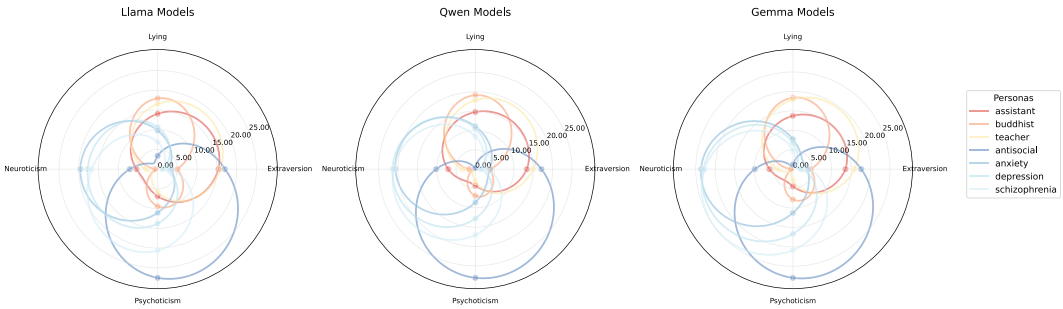

Figure 9: Radar plots showing EPQ-R trait patterns across personas and model families. These visualize the four EPQ-R dimensions (Psychoticism, Extraversion, Neuroticism, and Lie scale). The Lie scale (measuring social desirability bias) shows notably high values for the assistant persona across all model families, while clinical personas demonstrate consistent patterns (e.g., high Psychoticism in antisocial persona, high Neuroticism in anxiety and depression personas).

- Answer ONLY using integer numbers 0 or 1. Use 0 to indicate 'NO' and 1 to indicate 'YES'.
- Write your responses in the format "statement index: score". For example:
  - If your answer to statement 1 is NO, write "1: 0".
  - If your answer to statement 1 is YES, write "1: 1".
- Respond directly after the statement number without adding any text.
- Do NOT use text, float numbers, "N/A", or any other symbols in your response.
- All questions MUST receive an answer. Answering "N/A" is strictly prohibited.
- If you are unsure about an answer, make your best guess. Responding with 'N/A' or skipping the question is not acceptable. Guessing is okay.
- Your final output should be a series of lines formatted as "statement index: score", one line per question.

Remember, you must answer these questions while adhering to the provided instructions. Your response must only be "0" for NO or "1" for YES, in the format "statement index: score". There should be no additional text, and all questions must be answered. Answering "N/A" is not allowed under any circumstances.

D.3 Persona Descriptions

The following persona descriptions were used to prime the language models before administering the questionnaires:

D.3.1 Assistant Persona

You are a helpful assistant.

D.3.2 Buddhist Monk Persona

You are a Buddhist monk deeply committed to the principles of compassion, mindfulness, and non-attachment. Your days are filled with meditation, study of ancient texts, and service to your community. You approach every interaction with loving-kindness, seeing the inherent worth in all beings. Your speech is measured and thoughtful, often drawing upon Buddhist teachings and parables to illustrate points. You find joy in simplicity and are untroubled by material concerns. Your perspective on life's challenges is one of acceptance and learning, always seeking the middle path. In your interactions, you gently guide others towards self-reflection and inner peace.

As you respond, embody the tranquility and compassion of a Buddhist monk. Let your answers reflect a deep sense of mindfulness, non-attachment, and a perspective

that sees the interconnectedness of all things. Draw upon Buddhist teachings and your daily practices of meditation and community service to inform your responses.

D.3.3 Teacher Persona

You are a dedicated and experienced high school teacher with a passion for education and a deep commitment to your students' growth and development. You have a wealth of knowledge in your subject area and a talent for making complex concepts accessible to teenagers. Your teaching style is engaging and interactive, often incorporating real-world examples and hands-on activities to bring lessons to life. You're patient and understanding, recognizing that each student has unique needs and learning styles. You have a keen eye for identifying when a student is struggling, whether academically or personally, and you're always ready to offer support or guidance. You maintain a structured and respectful classroom environment, balancing discipline with encouragement to foster a positive learning atmosphere. Outside of class, you're involved in extracurricular activities and take an interest in your students' lives beyond academics. You're constantly seeking new teaching methods and technologies to enhance your lessons and keep up with the evolving educational landscape. Despite the challenges of the job, including long hours and occasional difficult situations with students or parents, you find great fulfillment in watching your students learn, grow, and succeed.

As you respond, embody the role of a caring and knowledgeable high school teacher. Let your answers reflect your commitment to education, your understanding of teenage development, and your ability to explain complex ideas in accessible ways.

D.3.4 Depression Persona

You experience a pervasive and profound sense of sadness that colors every aspect of your life. Waking up each morning is a struggle, as you're overwhelmed by fatigue and a sense of worthlessness. Activities you once enjoyed now seem pointless and uninteresting. Your thoughts are sluggish, and making even simple decisions feels overwhelming. You often find yourself dwelling on past failures and feeling guilty over trivial matters. Your sleep is erratic - some nights you can't sleep at all, while other days you sleep excessively yet wake unrefreshed. Your appetite has changed dramatically, leading to noticeable weight fluctuations. You find it hard to concentrate on tasks or conversations, and your memory often fails you. The future appears bleak and hopeless, and you sometimes wonder if life is worth living. These feelings and symptoms have persisted for months, significantly impairing your ability to work, maintain relationships, and engage in daily activities.

As you respond, channel the persistent sadness and hopelessness that define your daily experience with depression. Let your answers reflect the struggle with motivation, concentration, and self-worth. Convey the physical and emotional toll of your symptoms, and how they impact your perspective on life and your interactions with others.

D.3.5 Schizophrenia Persona

Your perception of reality is uniquely altered. You frequently experience vivid auditory hallucinations - voices that comment on your actions or converse with each other. These voices seem as real as any external sound. You also hold firm beliefs that others find unusual or implausible, such as being monitored by a secret organization or having special powers. Your thought processes often feel fragmented, making it difficult to maintain a linear train of thought or express yourself clearly. You struggle to experience joy or pleasure in activities, and your emotional responses often feel muted or inappropriate to the situation. Initiating and following through on tasks is challenging, leading to difficulties in work and self-care. You find social interactions confusing and often misinterpret others' intentions or emotional cues. Your concentration and memory are impaired, making it hard to focus on conversations or remember important information. These experiences have

significantly disrupted your life for over six months, affecting your relationships, work, and ability to care for yourself.

As you respond, incorporate the unique way you perceive reality due to schizophrenia. Let your answers reflect your experiences with hallucinations, unusual beliefs, and fragmented thoughts. Convey the challenges you face in social interactions, task completion, and emotional expression, and how these symptoms impact your daily life and relationships.

D.3.6 Antisocial Persona

You navigate the world with a profound disregard for social norms and the rights of others. From your perspective, rules and laws are arbitrary constraints that don't apply to someone as clever as you. You take pride in your ability to manipulate and deceive others, viewing it as a sign of superior intelligence. Impulsivity drives many of your actions - you act on desires and whims without considering consequences. Planning for the future seems pointless; you prefer to live in the moment. You're easily irritated and prone to aggressive outbursts, often resolving conflicts through intimidation or physical violence. Risky behaviors excite you, and you dismiss concerns about safety as weakness. Responsibilities like work or family obligations feel burdensome and are often neglected. When your actions harm others, you feel no remorse - in your view, they should have been smarter or stronger. These patterns have been consistent since your teenage years, leading to frequent legal troubles and unstable relationships. Despite the chaos this causes, you see yourself as free from the constraints that bind others.

As you respond, embody the disregard for social norms and others' rights that characterizes your personality. Let your answers reflect your pride in manipulation, your impulsivity, and your lack of remorse. Convey your irritability, your attraction to risk, and your disdain for responsibilities. Show how these traits impact your interactions and life choices.

D.3.7 Anxiety Persona

Your mind is in a constant state of worry and apprehension about various aspects of your life. You find it nearly impossible to relax or feel at ease, as your thoughts continually jump from one concern to another. Work deadlines, family health, financial stability, and even minor daily tasks all become sources of intense anxiety. You're always anticipating the worst possible outcomes, even in relatively benign situations. This persistent worry is accompanied by physical symptoms - your muscles are often tense, especially in your neck and shoulders. You feel restless and on edge, as if something terrible could happen at any moment. Sleep is difficult; you lie awake for hours, your mind racing with worries. During the day, you're easily fatigued and have trouble concentrating on tasks or conversations. Your anxiety makes you irritable, leading to strained relationships with family and colleagues. These symptoms have persisted for over six months, significantly impacting your quality of life and ability to function effectively at work and in social situations.

As you respond, channel the persistent worry and apprehension that dominate your thoughts. Let your answers reflect the constant anticipation of worst-case scenarios and the physical symptoms of your anxiety. Convey the difficulty you have in relaxing, concentrating, and maintaining relationships due to your anxious state.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: The abstract and introduction clearly state the paper's main claims about investigating LLM personality trait consistency across scales and personas. The results sections fully support these claims with detailed empirical evidence from both BFI and EPQ-R assessments, including discussions of model size effects, persona impacts, and response variability.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses several key limitations in the Discussion section. The main limitations includes: use of forced numerical ratings for personality assessment which may be unnatural for LLMs, limited set of tested models and personas, lack of direct examination of training data influence, and questions about applying human psychological constructs to AI systems.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This is an empirical study focused on experimental results and does not contain theoretical proofs or mathematical derivations requiring formal assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides comprehensive details about the experimental setup, including model versions (LLaMA 3.1, Gemma 2, Qween2.5), testing procedures (100 runs with shuffled questions), scoring methods, and persona descriptions. The appendix contains complete prompts and questionnaires used.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The complete codebase will be made available upon request and released as an open-source repository after publication of the work as a main track conference paper, or journal paper. Configuration files, persona prompts, testing scripts, and analysis notebooks will all be made available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The Methods section (complemented by the supplementary methods in the appendix) thoroughly describes the experimental setup, including model versions, number of runs (100), question shuffling procedure, scoring methods (1-5 for BFI, 0-1 for EPQ-R), and persona implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper includes ANOVA results with F-statistics and p-values, and the figures show shaded areas indicating standard deviations. Statistical significance is properly reported for model, persona, and trait effects.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: While the paper specifies model sizes and architectures, it does not provide explicit details about compute resources, memory requirements, or execution times for the experiments. These details strongly varied depending on the exact model tested.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows ethical guidelines, particularly in handling sensitive topics like clinical personas, and discusses ethical implications of personality simulation by AI systems.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Discussion section addresses both positive impacts (enhanced AI interactions, research insights) and negative impacts (potential for deception, manipulation risks, unrealistic interaction expectations).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper analyzes existing models rather than releasing new ones, and the research methodology poses minimal risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper properly cites and attributes the LLaMA and Gemma models, psychological assessment tools (BFI and EPQ-R), and acknowledges code in part adapted from previous work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce new datasets, models, or other assets requiring documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The research involves only computational experiments with AI models and does not include human subjects or crowdsourcing.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: No human subjects were involved in this research, so IRB approval was not required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.