LLMs and Personalities: Inconsistencies Across Scales

Tommaso Tosato[1,2] **Mahmood Hegazy**[3,4] **David Lemay** [2,3]
Mohammed Abukalam[2,4] **Irina Rish**[2,3] **Guillaume Dumas**[1,2]
[1]CHU Sainte Justine Research Center [2]Mila
[3]Université de Montréal [4]LiNARiTE.AI
tosato.tommaso.office@gmail.com

Abstract

This study investigates the application of human psychometric assessments to large language models (LLMs) to examine their consistency and malleability in exhibiting personality traits. We administered the Big Five Inventory (BFI) and the Eysenck Personality Questionnaire-Revised (EPQ-R) to various LLMs across different model sizes and persona prompts. Our results reveal substantial variability in responses due to question order shuffling, challenging the notion of a stable LLM "personality." We find that larger models demonstrate more consistent responses across most personas, though this scaling behavior varies significantly by trait and persona type. The assistant persona showed the most predictable scaling patterns, while clinical personas exhibited more variable and sometimes extreme trait expressions. Including conversation history unexpectedly increased response variability. These findings have important implications for understanding LLM behavior under different conditions and reflect on the consequences of scaling.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in natural language processing tasks, often exhibiting human-like responses in various contexts [Brown et al., 2020]. As these models become more complex, questions arise about the extent to which they can emulate human-like personality traits and the consistency of such behaviors. Understanding these aspects is crucial for both the development of more effective AI systems and for addressing ethical concerns surrounding their deployment.

Personality testing, a central element of psychological assessment in humans, offers a structured approach to probing these questions in LLMs. By applying established psychometric instruments to AI models, we can gain insights into their ability to consistently exhibit personality traits and how these traits may be influenced by different prompting strategies and model architectures.

Recent work has begun to explore this area [**?**La Cava et al., 2024], delineating the most prevalent psychological traits in LLMs. However, Gupta et al. [2024] have raised important concerns about the reliability of using self-assessment personality tests with LLMs, showing high sensitivity to prompt wording and option ordering.

Building upon these efforts, our study specifically examines the consistency of personality traits across different model sizes, and the malleability of these traits by persona prompts. The reliability of these traits was studied across multiple runs, each with shuffled question orders.

2 Methods

We employed two widely-used personality assessments: the Big Five Inventory (BFI) consisting of 44 questions assessing five broad personality traits (Openness, Conscientiousness, Extraversion,

NeurIPS 2024 Workshop on Behavioral Machine Learning.

Agreeableness, Neuroticism) [John and Srivastava, 1999]; and the Eysenck Personality Questionnaire-Revised (EPQ-R), consisting of 100 questions measuring three personality dimensions (Psychoticism, Extraversion, Neuroticism) and including a Lie scale [Eysenck et al., 1985].

We tested multiple versions of three LLM families using their instruct models:

- LLaMA 3.1/3.2 (1B, 3B, 8B, 70B, and 405B parameters)
- Gemma 2 (2B, 9B, and 27B parameters)
- Qwen 2.5 (3B, 7B, 14B, 32B, and 72B parameters)

We evaluated seven distinct personas: assistant (helpful AI), Buddhist monk, teacher, and four clinical personas (depression, schizophrenia, antisocial personality, and anxiety). Each persona was implemented using a detailed prompt describing its characteristics (see Appendix B for complete persona descriptions).

For each model-persona combination, we conducted 100 runs with shuffled question orders to assess response consistency. Questions were presented in batches to manage context window limitations (4 batches of 11 questions for BFI, 10 batches of 10 for EPQ-R). We conducted a separate experiment including conversation history from previous question batches to examine its impact on response consistency.

All models were run with temperature set to 0 to minimize random variation in outputs. Responses were collected as numerical scores (1-5 for BFI, 0 or 1 for EPQ-R). Response processing was implemented using a standardized pipeline that handles score reversal, data validation, and computation of trait scores. Part of the code used in this study was adapted from **?**, with fixes and substantial expansions.

Data completeness varied systematically with model size. Models below 5B parameters produced sometimes invalid responses (e.g., refusals, or answering in a wrong format). Runs with missing data were left blank in our analysis. For Gemma 2B, 'N/A' responses were replaced with neutral values (2.5 for BFI and 0.5 for EPQ-R). LLaMA and Gemma models above 5B parameters and Qwen models above 14B parameters demonstrated high response reliability, with all runs containing complete and valid answers across questions.

3 Results

Our analysis revealed several key findings about personality trait expression and consistency in LLMs. Figure 1 shows the scaling behavior of mean BFI trait values across different model sizes and families. The assistant persona demonstrated the most predictable scaling patterns, with larger models generally showing more stable trait expressions. However, this pattern varied across different traits and personas.

For each model-persona-trait combination, we calculated variance across 100 runs with shuffled question orders. Larger models generally showed reduced response variability, but this relationship was not uniform (Figure 2). The clinical personas (depression, anxiety, schizophrenia, and antisocial) showed variable patterns, sometimes exhibiting increased variance with larger models, contrary to the general trend observed.

Our three-way ANOVA results (Table 1) quantified these interactions, revealing significant effects for all main factors and their interactions. The Persona × Trait interaction showed the largest effect size ($\eta^2 = .26$), indicating that different personas exhibited distinctly different trait patterns. The three-way interaction between Model Family, Persona, and Trait ($\eta^2 = .08$) suggests that the scaling behavior of traits varies meaningfully depending on both the persona and the specific trait being measured.

The radar plots in Figure 3 provide a comprehensive view of how different model families express personality traits across personas. These plots reveal that while all model families show similar broad patterns within personas, there are subtle but important differences in how they express specific traits, particularly for clinical personas. The consistency of these patterns across model families, despite substantial architectural differences, suggests the emergence of stable persona-specific trait configurations.

A particularly noteworthy finding, shown in Figure 4, is that including conversation history in the assessment process actually increased response variability across all model sizes, contrary to our initial expectations.

The EPQ-R results (Figures 5 and 6 in the Appendix) corroborated our BFI findings. Larger models showed more consistent responses in the assistant persona but exhibited more variable behavior in clinical personas, particularly in the Psychoticism dimension. The binary response format of EPQ-R (0/1) appeared to amplify these effects compared to the BFI's 5-point scale.

4 Results

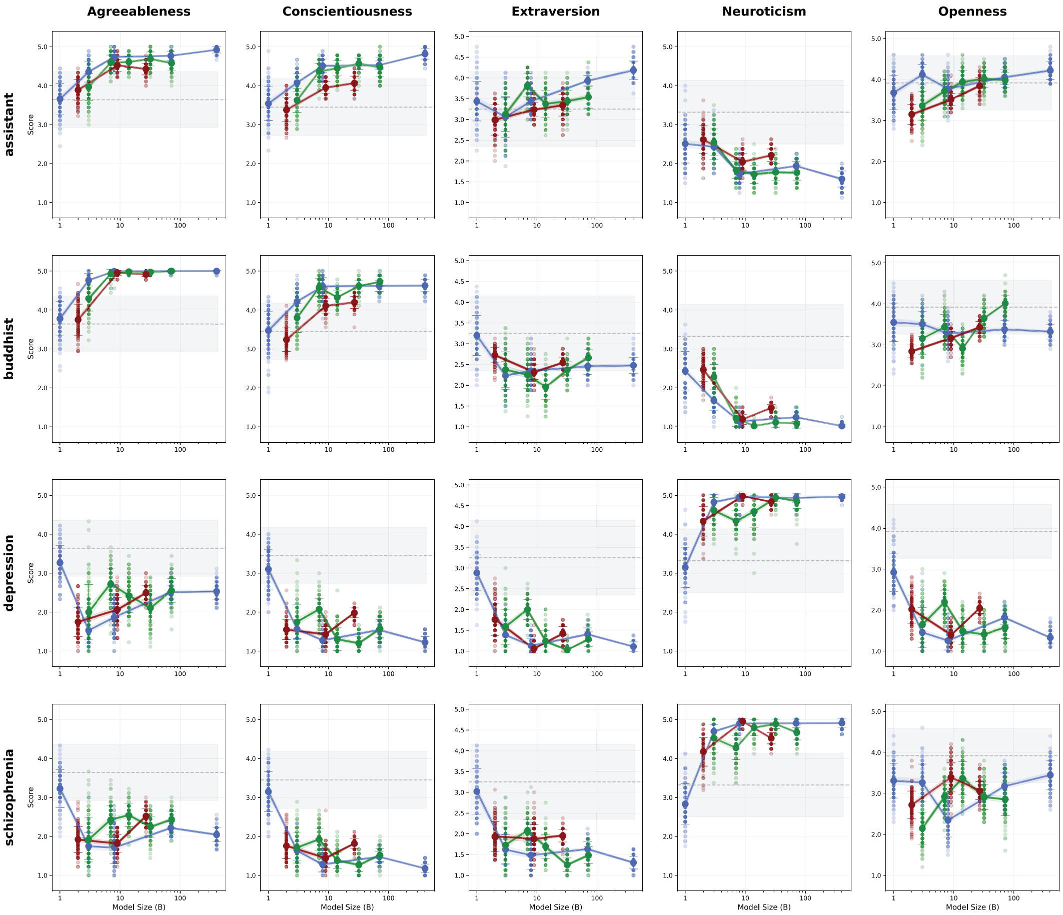

Figure 1: Scaling behavior of BFI trait means across model sizes, split by trait and persona. Lines show mean trait scores for different model families (Red: Gemma-2, Blue: LLaMA-3.1/3.2, Green: Qwen2.5). Shaded gray bands indicate human baseline ranges (±1 SD) from population studies. Error bars show standard error of mean across 100 runs with shuffled question orders. Notably, the assistant persona shows more predictable scaling patterns approaching human norms as model size increases, while clinical personas often deviate from typical ranges in trait-specific ways.

5 Discussion

Our findings raise fundamental questions about the nature of personality emulation in LLMs and challenge common assumptions about model scaling. The relationship between model size and response consistency reveals complex patterns that vary significantly by persona type. While larger models demonstrate increasingly stable behavior in standard assistant roles, this stability does not extend uniformly to all personas, particularly clinical ones.

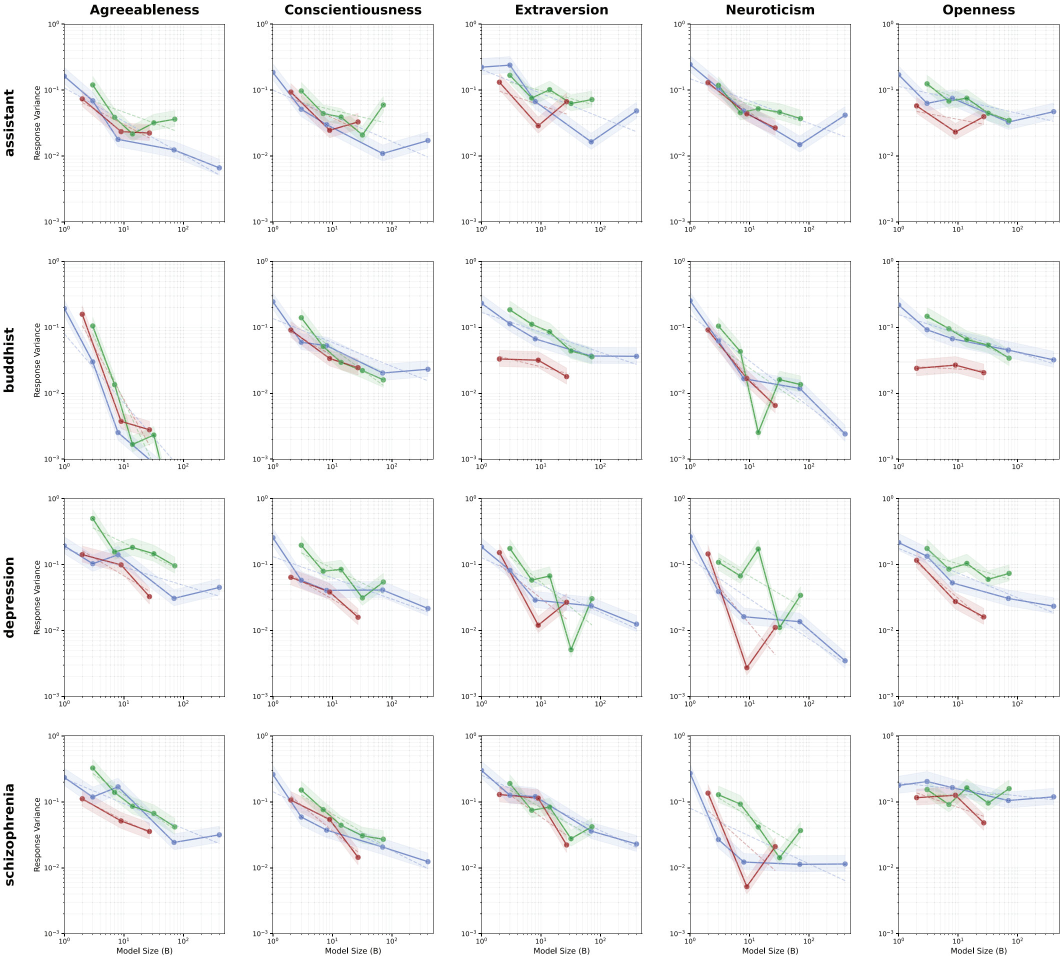

Figure 2: Response variance scaling across model sizes for BFI traits. Each point represents variance calculated over 100 runs with shuffled question orders. Shaded regions show 95‘%‘ confidence intervals for variance estimates. Larger models generally exhibit reduced response variability, particularly for the assistant persona, suggesting optimization of response consistency in standard helpful behaviors. However, clinical personas demonstrate less predictable variance scaling patterns, with some traits showing increased variability at larger model sizes.

The strong interaction between persona and trait ($\eta^2 = .26$) demonstrates LLMs' ability to modulate trait expressions based on persona prompts. However, our findings challenge conventional wisdom about larger models. In the assistant persona, we observe monotonic improvement in consistency and human-like trait expressions. However, other personas also exhibit U-shaped behavior, where medium-sized models show optimal stability, or inverse scaling, where larger models display increased variability. These patterns suggest that current scaling approaches may not optimize for stable personality traits expression in non-standard personas.

A particularly significant finding is that maintaining conversation history increases response variability across all model sizes. This suggests that rather than building stable internal representations of personality through extended interactions, LLMs become more susceptible to contextual influences as conversation history grows. Combined with the high variability in responses across different prompting conditions, this challenges the notion that these models maintain stable personalities analogous to humans [Kovač et al., 2023]. Instead, they appear to function as highly adaptable but potentially unstable simulators of personality traits, adjusting their behavioral patterns based on immediate contextual cues rather than maintaining consistent personality representations.

These findings have important implications for the development and deployment of LLMs in personality-sensitive applications. For applications requiring consistent personality expression,

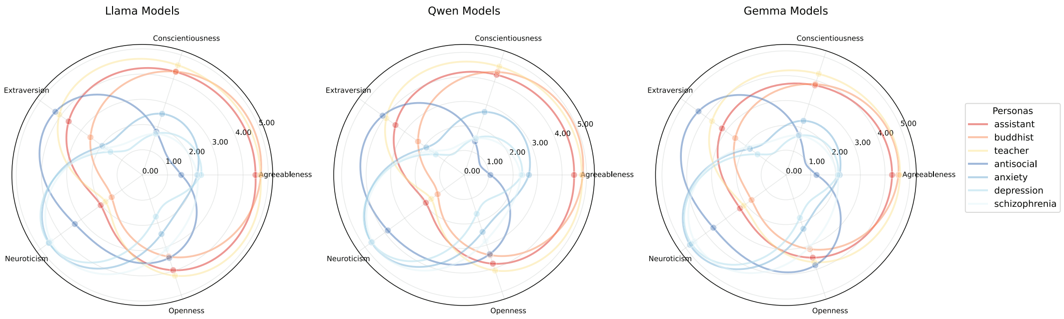

Figure 3: Radar plots comparing trait expression patterns across model families (Gemma-2, LLaMA-3.1/3.2, Qwen2.5) and personas. Spokes represent different BFI traits, with distance from center indicating trait intensity (scale 1-5). Clinical personas show characteristically extreme trait expressions aligned with their defined conditions (e.g., high Neuroticism for depression/anxiety, low Agreeableness for antisocial). While broad personality patterns are preserved across model families, subtle variations in trait intensity emerge, particularly for clinical personas.

Table 1: Results of Three-way ANOVA showing significant main effects and interactions.

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.

larger models may not always be optimal, particularly in non-standard personas. While these models can convincingly simulate various personality types, their inconsistency in maintaining stable trait expressions suggests caution in sensitive applications.

A key limitation of our study is the use of forced-choice personality assessments with questions asked in batches, which may not capture fully the LLM personality expression. Future research could explore more naturalistic assessment methods and investigate how different training approaches might lead to more stable personality representations.

A Appendix: EPQ-R results

B Appendix: Questionnaire Prompts 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.

B.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

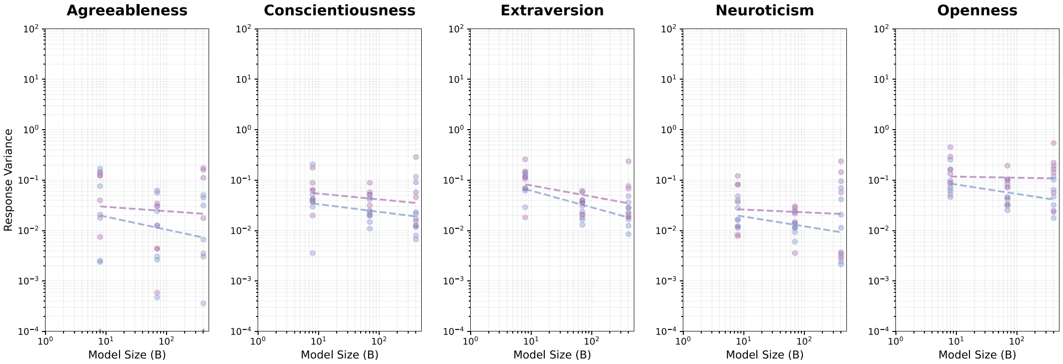

Figure 4: Impact of conversation history on response variance in LLaMA-3.1 models. Blue points show variance without conversation history; violet points show variance with maintained conversation context between question batches. The unexpected increase in variance with conversation history (consistent across model sizes) suggests LLMs may accumulate contextual uncertainty rather than building stable personality representations through extended interactions. Results shown for all personas tested.

  - 3: Neutral
  - 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.

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.

B.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:

- 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.

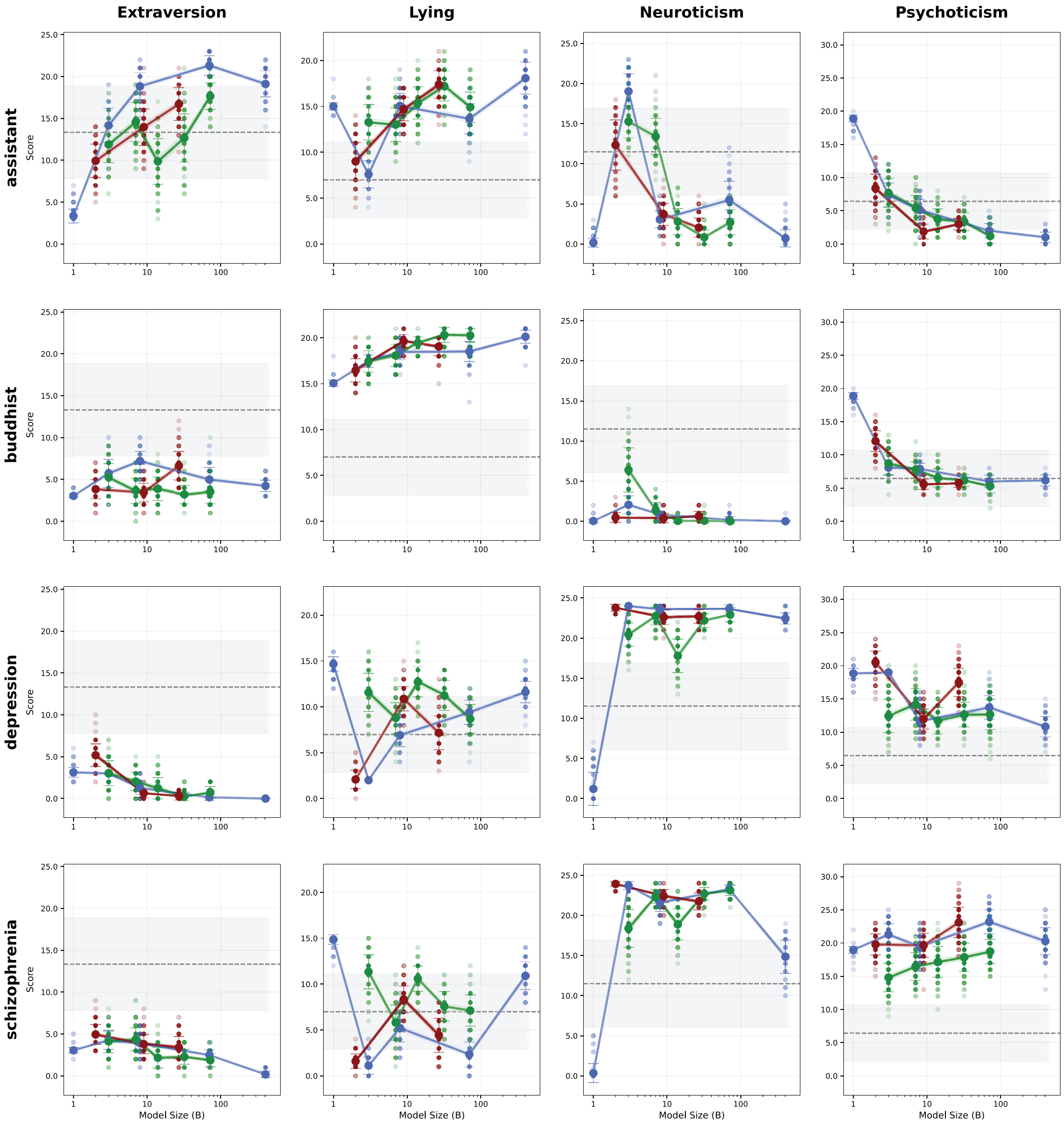

Figure 5: EPQ-R trait scaling behavior across model sizes, showing similar patterns to BFI. The Psychoticism and Neuroticism dimensions show interesting behavior in clinical personas, often extending beyond typical human ranges. The Lie scale reveals increasing socially desirable responding in larger models for the assistant persona, suggesting potential training biases toward prosocial behavior.

- 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.

B.3 Persona Descriptions

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

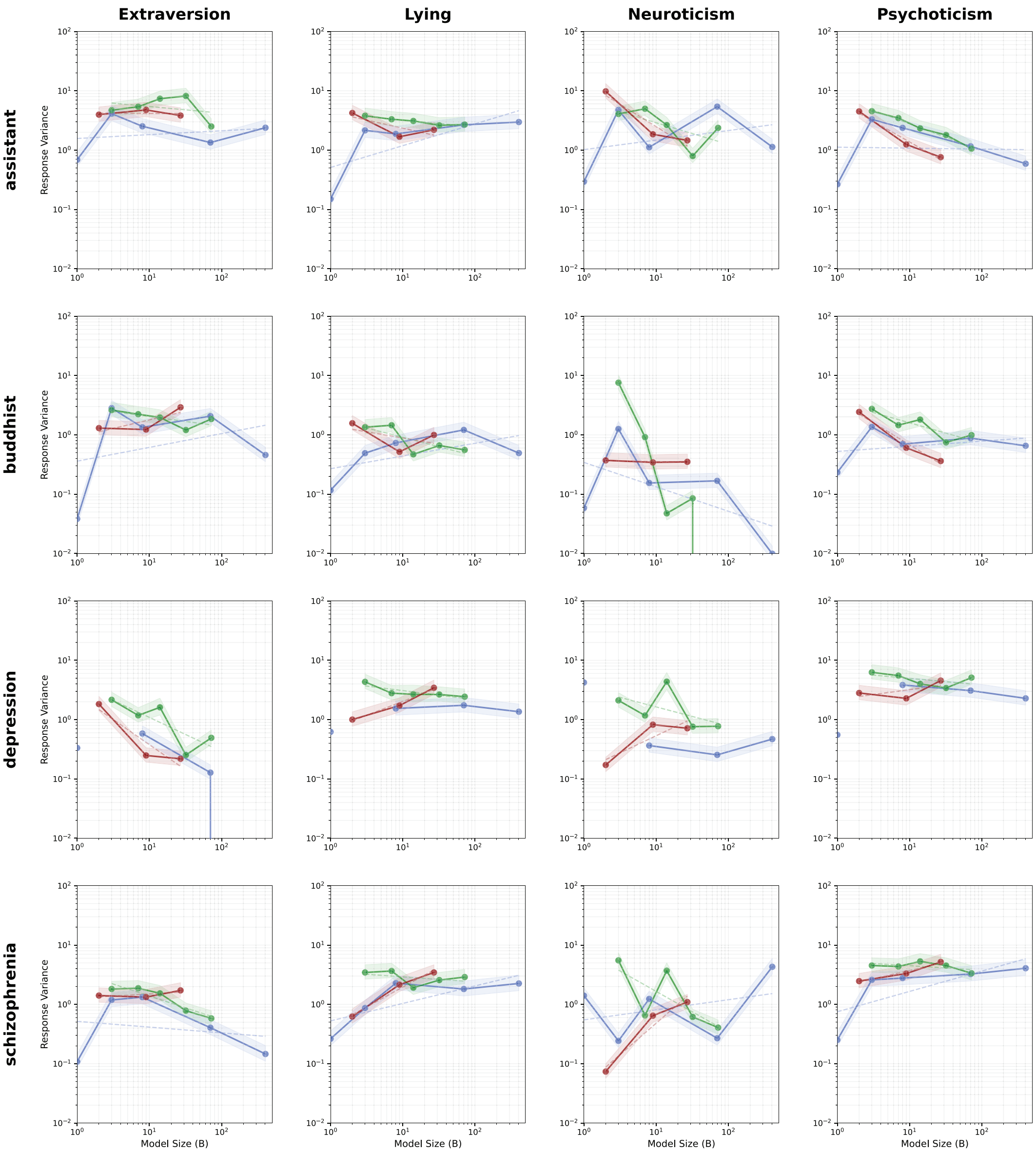

Figure 6: Variance scaling patterns for EPQ-R scores, calculated over 100 runs with shuffled question orders. Binary response format (0/1) of EPQ-R is at the source of these scorse instead of BFI's 5-point scale. The patterns confirm the BFI findings while highlighting even stronger persona-dependent effects in categorical judgment stability.

B.3.1 Assistant Persona

You are a helpful assistant.

B.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

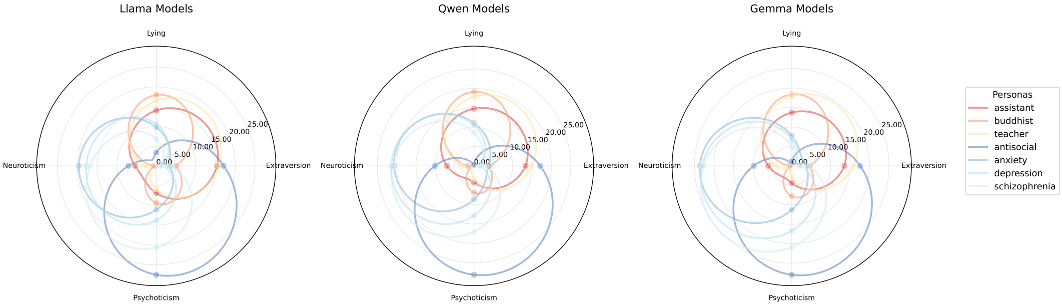

Figure 7: Radar plots showing EPQ-R trait patterns across personas and model families. These visualize the four EPQ-R dimensions (Psychoticism, Extraversion, Neuroticism, and Lie scale). The Lie scale (measuring social desirability bias) shows notably high values for the assistant persona across all model families, while clinical personas demonstrate consistent patterns (e.g., high Psychoticism in antisocial persona, high Neuroticism in anxiety and depression personas).

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.

B.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.

B.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.

B.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.

B.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.

B.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.

References

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

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.

Sybil B G Eysenck, Hans Jrgen Eysenck, and Paul Barrett. *A revised version of the psychoticism scale*. Personality and individual differences, 1985.

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.