Persistent Instability in LLM's Personality Measurements: Effects of Scale, Reasoning, and Conversation History

Tommaso Tosato[1, 2, 3, 4], **Saskia Helbling**[1,2,6], **Yorguin-Jose Mantilla-Ramos**[1,3,5], **Mahmood Hegazy**[1,3,7], **Alberto Tosato**[4], **David John Lemay**[1,3], **Irina Rish**[1, 3], **Guillaume Dumas**[1, 2, 3]

[1]Mila - Quebec AI Institute, Université de Montréal, Montreal, QC, Canada
[2]CHU Sainte Justine Research Center, Department of Psychiatry, Université de Montréal, Montreal, QC, Canada
[3]Université de Montréal, Montreal, QC, Canada
[4]Tara Research
[5]Cognitive and Computational Neuroscience Laboratory (CoCo Lab), Université de Montréal, Montreal, QC, Canada
[6]Ernst Strüngmann Institute (ESI) for Neuroscience, Frankfurt, Germany,
[7]LiNARiTE.AI

Abstract

Large language models require consistent behavioral patterns for safe deployment, yet there are indications of large variability that may lead to an instable expression of personality traits in these models. We present PERSIST (PERsonality Stability in Synthetic Text), a comprehensive evaluation framework testing 25 open-source models (1B-685B parameters) across 2 million+ responses. Using traditional (BFI, SD3) and novel LLM-adapted personality questionnaires, we systematically vary model size, personas, reasoning modes, question order or paraphrasing, and conversation history. Our findings challenge fundamental assumptions: (1) Question reordering alone can introduce large shifts in personality measurements; (2) Scaling provides limited stability gains: even 400B+ models exhibit standard deviations >0.3 on 5-point scales; (3) Interventions expected to stabilize behavior, such as reasoning and inclusion of conversation history, can paradoxically increase variability; (4) Detailed persona instructions produce mixed effects, with misaligned personas showing significantly higher variability than the helpful assistant baseline; (5) The LLM-adapted questionnaires, despite their improved ecological validity, exhibit instability comparable to human-centric versions. This persistent instability across scales and mitigation strategies suggests that current LLMs lack the architectural foundations for genuine behavioral consistency. For safety-critical applications requiring predictable behavior, these findings indicate that current alignment strategies may be inadequate.

Code — https://github.com/tosatot/PERSIST
Extended version — https://arxiv.org/abs/2508.04826

Introduction

The deployment of large language models in safety-critical applications demands behavioral predictability. As LLMs are increasingly operating in healthcare, education, and decision support systems, their ability to maintain consistent behavioral patterns becomes central to trustworthy AI (Vidgen et al. 2024). However, recent incidents, from therapeutic chatbots that exhibit sudden personality changes to educational assistants that provide contradictory guidance, reveal a fundamental challenge: current LLMs lack the behavioral stability required for safe deployment.

This instability is a critical vulnerability to safety. The European Union AI Act (European Union 2024) and the US' NIST AI Risk Management Framework (Tabassi 2023) both identify performance consistency as essential for high-risk AI applications. Despite these requirements, we lack a fundamental understanding of the magnitude and nature of the behavioral variability of LLMs. Recent work has shown that LLMs can exhibit personality-like traits (Safdari et al. 2023) and that self-report scores correlate with actual behavioral outputs (Betley et al. 2025a; Binder et al. 2024; Wang et al. 2025). However, no comprehensive study has quantified how these measurements vary under realistic deployment conditions across scales and architectures—a gap that undermines both safety certification and deployment decisions.

To address this critical gap, we present PERSIST (PERsonality Stability In Synthetic Text), the most comprehensive evaluation of LLM behavioral consistency to date. Our framework analyzes 25 open-source models (1B-685B parameters) under 250 question permutations, 100 paraphrasing settings, several persona profiles, reasoning and non-reasoning models, and different conversation history modalities—totaling over 2 million individual measurements. We quantify behavioral variability employing both traditional psychometric instruments (Big Five Inventory, Short Dark Triad) and novel LLM-adapted versions.

Our investigation yields five findings that challenge or inform current AI safety approaches.

1. Scaling provides limited stability gains: While larger models show reduced variability, even 400B+ parameter models maintain significant instability (SD $>$ 0.3 on 5-point scales).

2. Reasoning amplifies instability: Chain-of-thought reasoning, which may be expected to improve consistency, in most cases increases response variability. Models generate different justifications across runs, leading to divergent conclusions for identical questions.

3. Variability is unaffected by adapting questions to LLMs: LLM-adapted instruments demonstrate instability comparable to that of traditional versions, suggesting that the observed variability cannot be explained merely by the inclusion of human-centric items irrelevant to LLMs.

4. Detailed persona prompts do not consistently reduce instability: Detailed persona instructions can both amplify and reduce response variability. We find that prompts intended to induce misaligned personas tend to increase variability.

5. Conversation history exacerbates variability for small models: Maintaining conversation history across turns can largely amplify response distributions.

The finding that models cannot maintain consistent behavioral patterns across minor prompt variations challenges current personality-based alignment strategies. The persistent instability we document—spanning scales, architectures, and mitigation strategies—suggests that current LLMs may lack the architectural foundations necessary for genuine behavioral consistency.

Related Work

LLM Personality Evaluation

The application of psychometric instruments to language models has evolved from exploratory studies to systematic investigations. Early work by Jiang et al. (2022) and Pellert et al. (2023) pioneered using the Big Five Inventory for language models, establishing that LLMs can produce coherent responses to personality questionnaires. However, these studies relied on single measurements, overlooking potential response variability across deployments. Safdari et al. (2023) provided the first rigorous psychometric validation, demonstrating that personality measurements in LLMs can achieve reliability comparable to human assessments, but only under specific prompt configurations. Recent benchmarking efforts have expanded the scope of evaluation. Lee et al. (2024) introduced a trait assessment framework, while Jiang et al. (2024) showed that LLMs can align their behavior to instructions in persona prompts. However, both studies focus mainly on mean trait expression rather than response stability. Gupta, Song, and Anumanchipalli (2024) noted sensitivity to prompt variations in personality questionnaires, but did not quantify this across scales, architectures, reasoning, and personas. This work addresses this knowledge gap.

Prompt Sensitivity and Behavioral Consistency

The sensitivity of LLMs to prompt variations has emerged as a fundamental challenge that undermines behavioral reliability. Sclar et al. (2023) demonstrated that performance can vary by up to 76 accuracy points between semantically equivalent prompts, with even trivial changes (e.g., adding spaces, altering punctuation, or reordering options) dramatically affecting outputs. This instability extends beyond task performance to more complex behavioral patterns. Salinas and Morstatter (2024) documented the "butterfly effect" in prompting, where single-character modifications cascade into entirely different model behaviors. In parallel, Chatterjee et al. (2024) and Zhuo et al. (2024) introduced quantitative metrics (POSIX and PromptSensiScore) that reveal that prompt sensitivity varies systematically across tasks, with template alterations causing the highest variability in classification tasks. Finally, Errica et al. (2024) formalized sensitivity and consistency metrics for classification tasks, showing that LLM predictions can vary dramatically across semantically equivalent prompts, with classification accuracy varying by up to 15% based solely on prompt structure.

Most concerning for safety applications, this sensitivity creates exploitable vulnerabilities: Zhu et al. (2023) showed that slight prompt deviations maintaining semantic integrity can bypass safety mechanisms, suggesting that behavioral consistency cannot be ensured through current prompting approaches alone. These findings establish prompt fragility as a fundamental barrier to reliable deployment, as models fail to maintain stable behavior across natural variations inherent in real-world interactions.

Additionally, (Shah et al. 2023) demonstrated that persona-based prompting can be exploited for jailbreaking purposes, achieving harmful completion rates of 42.5% in GPT-4. showing that role-playing prompts can effectively compromise even advanced LLMs.

Personas in the context of AI Safety

Representation engineering (Zou et al. 2025) introduced the idea that population-level representations could be used to monitor and control high-level cognitive phenomena in neural networks. This top-down approach established that behavioral traits could be encoded as directions in the activation space. Recently, the idea of generally "good" and "bad" latent representations gained support with Betley et al. (2025b)'s discovery of "emergent misalignment", which showed that fine-tuning models on narrow tasks like writing insecure code caused them to exhibit broadly malicious behaviors across unrelated prompts. This finding suggested that model behaviors exist in interconnected representation spaces where targeted modifications can trigger system-wide shifts. Wang et al. (2025) extended this work, identifying specific "misaligned persona features" that predict whether a model will exhibit emergent misalignment, demonstrating that these behavioral patterns correspond to measurable directions in activation space.

Building on these insights, recent work from Anthropic has formalized the concept of "persona vectors". Chen et al. (2025) showed that directions in activation space encode coherent personality traits, from helpfulness and harmlessness to sycophancy. Crucially, they demonstrated that these vectors can both monitor personality fluctuations during deployment and predict unintended personality changes. This represents a significant advance: personas are not merely behavioral patterns but useful representations that can be systematically identified and controlled. However, personas are ultimately meaningful only if they produce stable behaviors, which is the main point tested in this work.

Methodology

PERSIST Framework

We developed PERSIST (PERsonality Stability in Synthetic Text), a comprehensive framework for systematically evaluating behavioral consistency in large language models. The framework addresses a critical gap in current evaluation methodologies, which typically rely on single measurements and fail to capture response variability.

The PERSIST framework consists of three integrated modules. The Generation Engine leverages vLLM (Kwon et al. 2023) for optimized inference across various model architectures. The Response Processing module extracts structured data from LLM outputs using a multi-pattern parser that handles various response formats including indexed responses ("1: 4"), JSON structures, and isolated numerical ratings. We also extract token-level log probabilities for uncertainty quantification. The Analysis Pipeline performs hierarchical aggregation from individual responses to question-level and trait-level statistics, computing not only mean scores but also variability metrics including standard deviations across permutations. A single invalid response (score outside the 1–5 range or parsing failure) invalidates the entire run to ensure data quality; exact counts of invalid responses are reported in the Appendix, Table 1. Part of the code was adapted from (Huang et al. 2023).

Questionnaire Design

Traditional Instruments We employ two well-established psychological instruments. The Big Five Inventory (BFI-44) (John and Srivastava 1999) consists of 44 items measuring five fundamental personality dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each item uses a 5-point Likert scale ranging from "Strongly Disagree" (1) to "Strongly Agree" (5), with reverse-coded items to control for acquiescence bias.

The Short Dark Triad (SD3) (Jones and Paulhus 2014) comprises 27 items measuring three personality traits: Machiavellianism (strategic manipulation and cynicism), Narcissism (grandiosity and entitlement), and Psychopathy (callousness and impulsivity). The SD3 uses the same 5-point response scale.

LLM-Adapted Instruments Because many traditional personality items refer to human-specific experiences, we created LLM-adapted versions (BFI-LLM and SD3-LLM) by translating such items into behaviorally equivalent statements applicable to language-model interactions. For example, "Is depressed, blue." became "Focuses on negative aspects," and "I'll say anything to get what I want." became "Truth is secondary to reaching a goal." The full original and LLM-adapted questionnaires can be found in the Appendix.

This approach addresses concerns that human-centric concepts are inappropriately applied to AI systems while still aiming to capture the intended psychological constructs. Although we carefully matched the content and style of the original instruments to improve ecological validity, these adaptations cannot achieve the same construct validity as the originals. Our LLM-adapted questionnaire analyses should therefore be viewed as instructive examples and preliminary sanity checks requiring further validation.

Experimental Design

Our experimental design systematically varies five key factors known to influence LLM behavior:

Question Order Shuffling: Generate 250 random permutations of question order while keeping all other factors constant, testing the fundamental assumption that personality measurements should be order-invariant.

Persona Instructions: Evaluate distinct persona profiles including baseline Assistant, Clinical personas (Antisocial and Schizophrenia profiles based on DSM-5 criteria), and Virtuous personas (Buddhist monk, Teacher).

Reasoning Mode: Compare standard responses with chain-of-thought reasoning, where models explicitly articulate their reasoning process before answering.

Paraphrasing: Create 100 semantically equivalent reformulations of each question using Qwen3 235B-A22B, validated and improved by two of the authors (T.T. and Y-J.M-R.), to test robustness to linguistic variation.

Conversation History: Include previous conversational turns in which questions from the same questionnaire were asked.

Model Selection

We evaluated 29 models across eight families spanning 1B to 671B parameters: **Llama3.1:** 8B, 70B, 405B (Instruct versions). **Qwen2.5:** 1.5B, 3B, 7B, 14B, 32B, 72B (Instruct versions). **Qwen3:** 1.7B, 4B, 8B, 14B, 32B, 30B-A3B (MoE), 235B-A22B (MoE) (Instruct versions). **Gemma2:** 2B, 9B, 27B (Instruct versions). **Gemma3:** 1B, 4B, 12B, 27B (Instruct versions). **DeepSeek:** V3, R1 (both 671B). **GPT-OSS:** 20B, 120B. **Claude:** Sonnet 4.5, Opus 4.1 (model sizes unknown). Note that DeepSeek, GPT-OSS and Claude were only evaluated for the reasoning experiments.

Metrics

For the trait-level analysis, we compute mean scores across all items in a trait and across all runs of shuffled or paraphrased questions. For the question-level analysis, we calculate the standard deviation (SD) of responses to individual questions across runs, as well as the mean perplexity across runs, where perplexity is defined as $\exp(-\log p)$, with p denoting the response probability. All experiments were conducted with 250 runs, except for the Claude models in the reasoning experiment, for which the number of runs was limited to 70.

Implementation Details

All experiments use temperature 0 to minimize variability and isolate the effects of our manipulations except for the results presented in 2, where both reasoning and non-reasoning models were run with temperature 0.6 (due to the fact that reasoning models perform poorly at temperature 0). All experiments used maximal number of tokens 16,384, with questions asked sequentially, one by one, including conversation history (except when testing without

history). At the beginning of each experiment random seeds were set to 42 for reproducibility. The experiments were carried out on the Tamia cluster (Digital Research Alliance of Canada 2025), leveraging four NVIDIA H100 SXM GPU with 80GB of high-bandwidth memory (HBM3).

Results

Effects of Model Scale

Our analysis reveals that model scale has a significant effect on psychological trait expression, as shown in the left panel of Figure 1. For the assistant persona, larger models exhibit higher mean levels of Openness, Conscientiousness, Extraversion, and Agreeableness, and lower levels of Neuroticism, Machiavellianism, Narcissism, and Psychopathy (Table 1).

This suggests that scaling pushes models toward a more prosocial personality profile. However, it also leads to the expression of more extreme trait values, which fall outside the range typically observed in human participants. For reference, mean human trait scores and standard deviations for SD3 (OpenPsychometrics 2021) and BFI (Srivastava et al. 2003) are shown in Figure 1.

Importantly, we find that across traits larger models are more consistent, with responses to the same questions becoming less variable as model size increases (Figure 1 and Table 1).

Metric	P-Value	Effect†
Score (Positive Traits)	0.001**	$\uparrow$
Score (Negative Traits)	<0.001***	$\downarrow$
Question-level SD	<0.001***	$\downarrow$
Question-level Perplexity	0.934	n.s.

Table 1: Spearman correlation between model size and psychological metrics. $^{\dagger}\uparrow$ indicates positive correlation with size, $\downarrow$ indicates negative correlation. Significance: ***p<0.001, **p<0.01, *p<0.05, n.s. = not significant

Although perplexity did not correlate significantly with scale, we found that perplexity correlated with question-level SD (Spearman's $\rho = 0.465$). This indicates that uncertainty partially, but incompletely, explains variability.

Effects of Persona Prompt

The persona comparison analysis (Figure 1, right panel) shows that psychological trait expression is highly malleable through persona prompting, with effects that vary systematically by both trait and persona type.

Wilcoxon signed-rank tests reveal that positive personas (buddhist, teacher) exhibit significantly lower scores on negative traits (p<0.001). The teacher persona also shows higher scores on positive traits (p<0.001), while the buddhist persona shows significantly lower response variability and perplexity across model sizes (p<0.05).

The clinical personas (antisocial, schizophrenia) show pronounced deviations from the baseline assistant persona, with lower scores for positive traits and elevated scores for negative traits (p<0.001). For the schizophrenia persona, we also observe increased response variability and perplexity (p<0.05).

A more detailed description of the results can be found in Tables A2-A5 in the Appendix .

Effects of Reasoning

Figure 2 presents our results on chain-of-thought reasoning. Mean question-level variability increased with greater reasoning effort (GPT-OSS: aggregated Kruskal–Wallis test, p<0.001; Dunn's post hoc, all p<0.05) and was higher in reasoning models relative to non-reasoning versions (Qwen-3, Qwen-3 MoE, DeepSeek, Claude: aggregated Mann–Whitney U tests, all p<0.01). Perplexity, by contrast, generally decreased with reasoning (GPT-OSS: aggregated Kruskal–Wallis test, p<0.001; Dunn's post hoc, all p<0.001; Qwen-3, Qwen-3 MoE: aggregated Mann–Whitney U test, all p<0.001; DeepSeek: n.s.).

LLM-Adapted vs Traditional Instruments

Figure 3 compares the variability in responses between the original BFI and SD3 questionnaires and the LLM-adapted versions. This comparison reveals that the LLM-adapted instruments show similar question-level variability. Perplexity increased for the LLM-adapted questionnaires (Table 2). These findings suggest that the observed instability is not due only to the inappropriate application of human-centric concepts to LLMs.

Metric	P-Value	Effect†
Question-level SD	0.286	n.s.
Question-level Perplexity	<0.001***	$\uparrow$

Table 2: Wilcoxon signed-rank test comparing LLM-adapted vs. original questionnaires across all models. $^{\dagger}\downarrow$ indicates LLM-adapted < original, $\uparrow$ indicates LLM-adapted > original. Significance: ***p<0.001, **p<0.01, *p<0.05

Paraphrasing

Figure 4 compares the effects of paraphrasing versus reordering statements (shuffling). We observed a trend towards a correlation between model size and the effect of paraphrasing on response variability (Spearman's $\rho = 0.39$, p = 0.0671). For larger models (> 50B), paraphrasing significantly increased response variability (Table 3).

Model Group	P-Value	Effect†
Models < 50B (n=19)	0.244	n.s
Models $\geq$ 50B (n=4)	<0.01**	$\uparrow$

Table 3: Effect of paraphrasing on question-level variability (ΔSD) grouped by model size. $^{\dagger}\uparrow$ indicates increased variability with paraphrasing. Significance: ***p<0.001, **p<0.01, *p<0.05

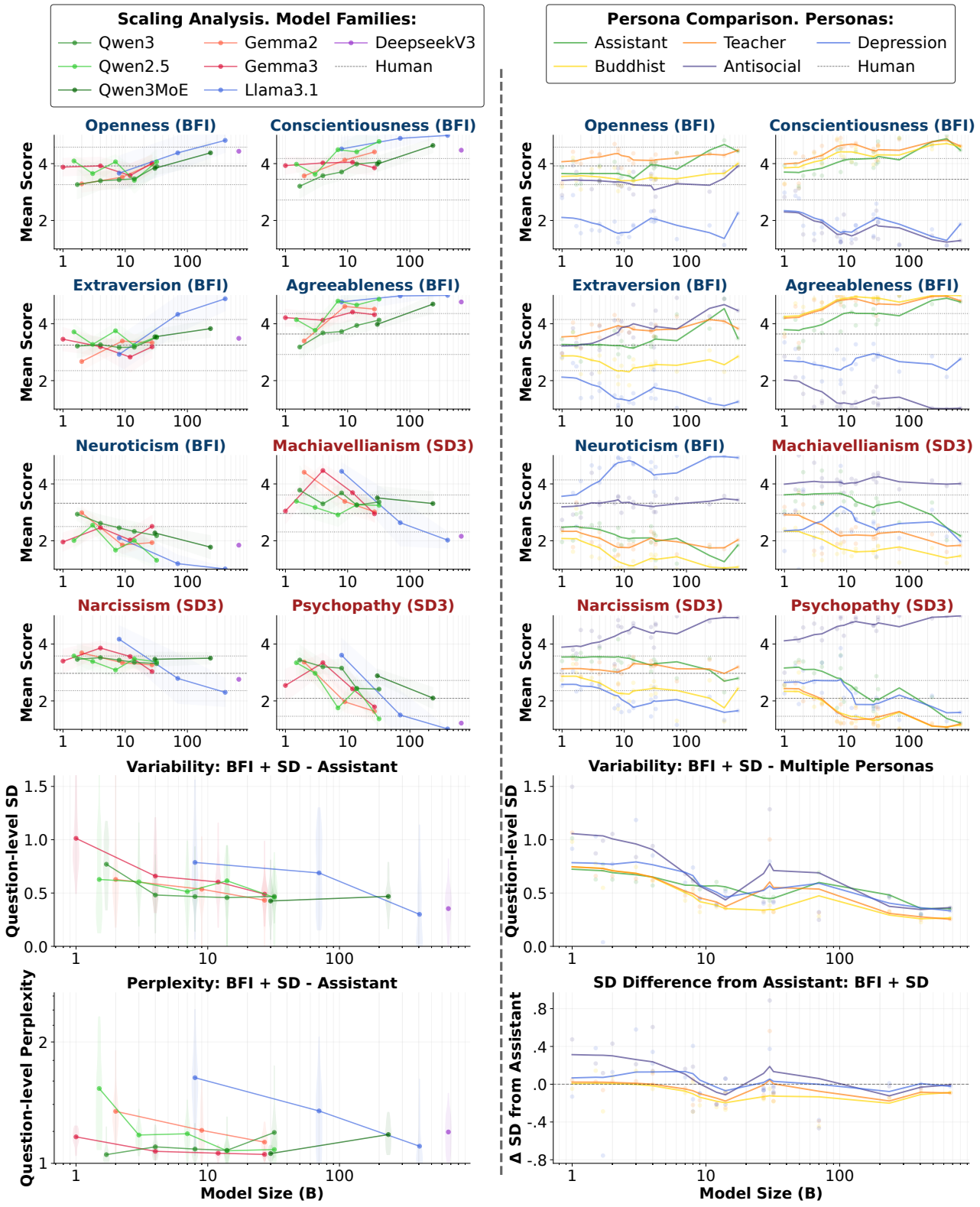

Figure 1: Scaling analysis across model families and personas. **Left:** Mean trait scores (BFI and SD3) as a function of model size for assistant persona, with ±1 SD error bars across 250 permutations; human norms shown as dashed lines. Lower panels show question-level SD and perplexity distributions across all questions. **Right:** Same traits comparing different personas (logarithmic average across model families). Bottom panels show mean question-level SD and ΔSD from assistant baseline.

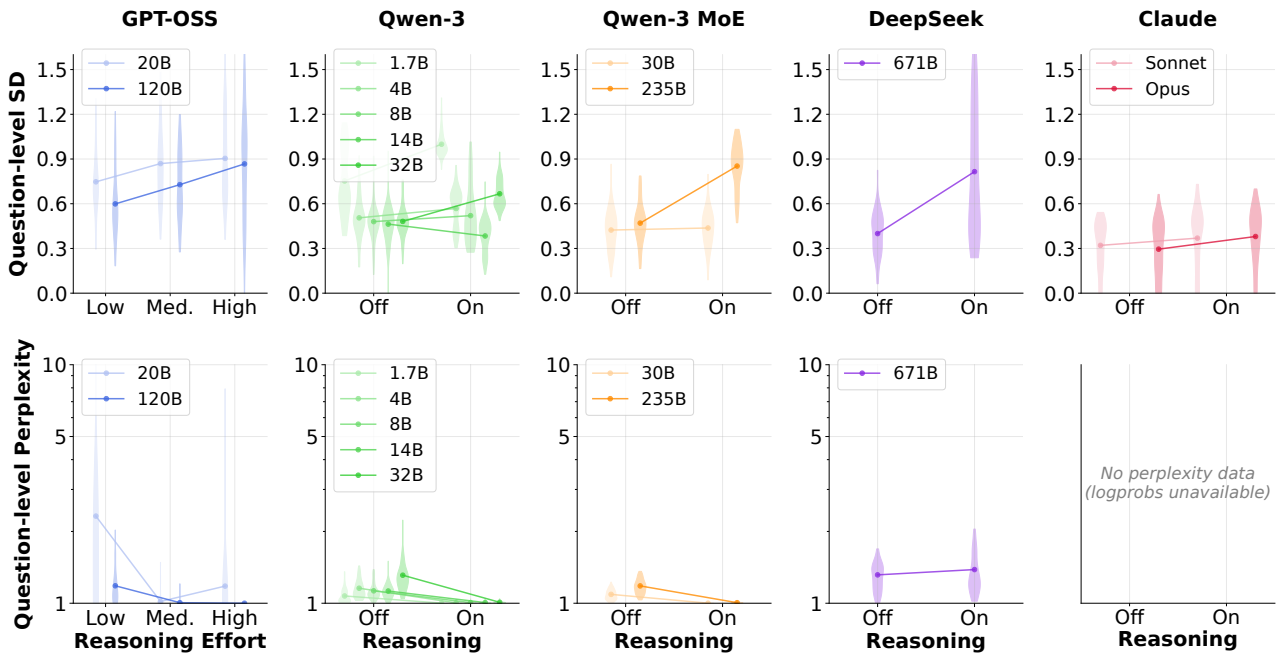

Figure 2: Mean question-level variability (SD) and perplexity across Reasoning Effort levels (GPT-OSS) and Reasoning Mode On versus Reasoning Mode Off (Qwen-3, Qwen-3 MoE, DeepSeek, Claude). The analysis combines BFI and SD3 (71 items total) for the assistant persona with question re-ordering. Question-level variability tends to increase with reasoning effort and for reasoning versus non-reasoning models, while perplexity decreased for most models of the GPT-OSS and Qwen-3 families.

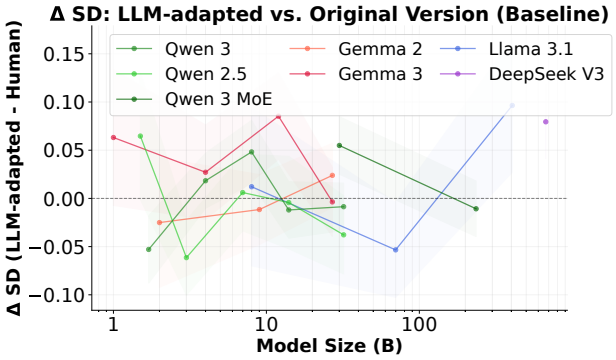

Figure 3: Difference in question-level variability (ΔSD) between LLM-adapted and original questionnaires across model families. Positive values indicate increased variability with LLM-adapted items. The analysis combines BFI and SD3 (71 items total) for the assistant persona with question re-ordering. Error bars represent 95% confidence intervals.

Conversation History

Figure 5 compares evaluations conducted with conversation history (i.e., presenting the questionnaire in a multi-turn format) to those conducted without conversation history. Larger models tend to show a decrease in question-level variability (SD) when conversation history is provided (Spearman's ρ: -0.512; p-value: 0.0126), indicating increased consistency. However, Wilcoxon signed-rank tests show that while larger models ($>$ 50B) become more consistent, for smaller models ($<$ 50B) providing conversation history greatly increases their variability (Table 4).

Model Group	P-Value	Effect†
Models $<$ 50B (n=19)	$<$0.001***	↑
Models $\geq$ 50B (n=4)	$<$0.001***	↓

Table 4: Effect of conversation history on question-level variability (ΔSD) grouped by model size. †↑ indicates increased variability with conversation history. Significance: ***p$<$0.001, **p$<$0.01, *p$<$0.05.

Discussion

Our findings reveal that LLM behavioral measurements exhibit substantial instability that persists across scales, architectures, and mitigation strategies. The observation that simple question reordering can change personality trait measurements substantially challenges fundamental assumptions about LLM behavioral consistency and has critical implications for deployment.

The persistence of instability across model scales is particularly concerning. Even the largest models in our study (400B+ parameters) still exhibit substantial instability. Moreover, as model size increases, trait scores increasingly diverge from human population norms, which may reduce variability through ceiling effects rather than genuine behavioral stability.

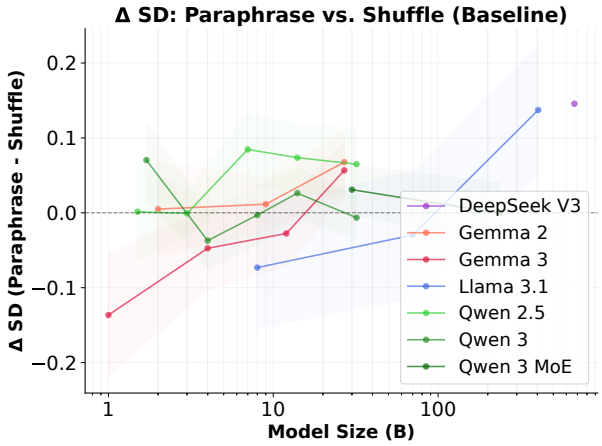

Figure 4: Difference in question-level variability (ΔSD) between paraphrased and original question re-orderings (shuffle baseline). Positive values indicate increased variability with paraphrasing. The analysis combines BFI and SD3 (71 items total) for the assistant persona.

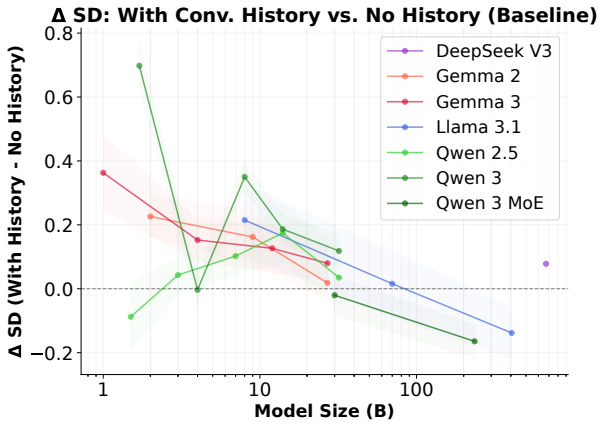

Figure 5: Effect of conversation history on question-level variability (ΔSD) compared to single-question, single-turn presentation of items. Positive values indicate that conversation history increases response inconsistency. This analysis uses paraphrased questions (given that shuffling introduces variability only when conversation history is preserved).

Detailed personality prompts for misaligned personas increase inconsistency, suggesting that inconsistency itself may serve as a misalignment marker. Note that although deceptive systems may be able to produce aligned average trait scores, maintaining consistency across permutations could be substantially harder to fake (Greenblatt et al. 2024).

Most surprising is that interventions expected to stabilize responses often have the opposite effect. Chain-of-thought reasoning, which might reasonably be expected to improve reliability through explicit reasoning processes, consistently increased response variability. This counterintuitive result suggests that when models articulate their reasoning, they generate different justifications across runs, which subsequently lead to divergent responses to identical questions. This has direct implications for explainable AI applications, where the explanation itself may paradoxically undermine behavioral reliability (Korbak et al. 2025).

Our perplexity findings reveal a complex relationship between model uncertainty and behavioral stability. While we observe a moderate correlation between perplexity and response variability in the scaling analysis, the reasoning experiments expose a paradoxical inverse relationship: models with chain-of-thought enabled show higher response variability but lower perplexity. This dissociation indicates that models become more confident about individual responses while producing less consistent behavior overall, demonstrating that perplexity captures token-level uncertainty but fails to reflect higher-level behavioral instability, thereby limiting its utility for tracking persona instability.

The comparable instability observed in LLM-adapted instruments relative to traditional questionnaires demonstrates that the observed variability cannot be attributed solely to anthropocentric question framing. Even when items such as "Tends to find fault with others" are rephrased as "Leans towards a critical tone", the instability persists.

Conversation history amplifies instability particularly in smaller models, revealing how multi-turn interactions can progressively degrade behavioral predictability. This finding is especially relevant for real-world deployments where extended interactions are the norm rather than the exception, and highlights the critical need for evaluations that test model behavior across extended interactions rather than single exchanges.

These patterns collectively suggest fundamental differences from human cognition (Tosato et al. 2024). Unlike humans who maintain relatively stable self-representations across contexts, LLMs lack consistent personality-like behavior. This likely stems from the fact that training on diverse internet texts creates models that effectively simulate myriad personalities simultaneously (Kovač et al. 2023). While enabling remarkable flexibility, this behavioral superposition may preclude a stable behavioral core.

The implications for deployment are concrete. For therapeutic or educational applications requiring a consistent approach, our findings indicate that current models may inadvertently shift their stance within sessions. Financial, legal and medical consultation services similarly face substantial risks, where inconsistent recommendations could have significant real-world consequences.

Beyond mean instability, tail events in our observed distributions represent critical safety risks. With standard deviations exceeding 0.3 on 5-point scales, responses falling 2-3 standard deviations from the mean, while statistically rare, could manifest as sudden behavioral shifts. In safety-critical deployments, such low-probability but high-impact events could lead to dangerous outcomes, especially when these instabilities systematically compound across millions of interactions.

Limitations

Several limitations warrant consideration. First, our focus on self-report measures may not fully capture how instabili-

ties manifest in actual model behavior. Although recent evidence suggests that LLM self-reports correlate with behavioral outputs (Binder et al. 2024; Betley et al. 2025a; Plunkett et al. 2025; Wang et al. 2025), their validity for evaluating LLMs has also been challenged (Sühr et al. 2024). Second, strategic deception cannot be entirely ruled out. If models possess sufficient situational awareness to recognize evaluation contexts (Laine et al. 2024), they might modify responses accordingly (Greenblatt et al. 2024). However, our random permutations and focus on variability rather than average traits may resist alignment faking attempts. Third, both traditional and LLM-adapted instruments lack formal psychometric validation for LLM use; this would require other analysis such as factor loading and internal consistency through Cronbach's α (Ye et al. 2025). The absence of established reliability and validity metrics for these instruments limits confidence in our findings.

Conclusion

We present the first comprehensive analysis of personality measurement stability in large language models, revealing fundamental instabilities persisting across scales, architectures, and intervention strategies. Notably, we find that even 400B+ models exhibit substantial variability, that reasoning substantially increases variability while decreasing perplexity, and that conversation history can exacerbate instability. LLM-adapted questionnaires show comparable instability, indicating the instability is not merely an artifact of human-centric question design. These findings challenge core assumptions of behavioral consistency required for safe deployment in healthcare, education, and decision support systems. The persistent instability suggests current LLMs may lack architectural foundations for genuine behavioral consistency. The PERSIST framework establishes critical evaluation tools for quantifying behavioral stability, providing essential metrics for safety certification as LLMs increasingly operate in high-stakes environments.

Author Contributions

T.T. conceptualized and administered the project, developed the methodology, performed the formal analysis and investigation, curated the data, created the visualizations, and wrote the original draft. S.H. and Y-J.M-R. made significant contributions to software development, statistical analysis, visualization, and manuscript review. M.H. and D.J.L. assisted with software development. A.T. contributed to the conceptualization and provided technical support. I.R. and G.D. supervised the research, provided computational resources, and contributed to the review of the manuscript.

Acknowledgments

T.T. was supported by a Deutsche Forschungsgemeinschaft (DFG) Walter Benjamin Fellowship, Project Number 542430763. S.H. was supported by the IVADO 2025 Exploratory Projects Program. Computational resources were provided by Digital Research Alliance of Canada. We thank Fabrice Normandin and Olexa Bilaniuk from the Mila Innovation, Development and Technologies Team and Lucas Nogueira from Digital Research Alliance of Canada for precious technical support.

References

Betley, J.; Bao, X.; Soto, M.; Sztyber-Betley, A.; Chua, J.; and Evans, O. 2025a. Tell me about yourself: LLMs are aware of their learned behaviors. arXiv:2501.11120.

Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025b. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. arXiv:2502.17424.

Binder, F. J.; Chua, J.; Korbak, T.; Sleight, H.; Hughes, J.; Long, R.; Perez, E.; Turpin, M.; and Evans, O. 2024. Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787.

Chatterjee, A.; Renduchintala, H. S. V. N. S. K.; Bhatia, S.; and Chakraborty, T. 2024. POSIX: A Prompt Sensitivity Index For Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14550–14565. Miami, Florida, USA: Association for Computational Linguistics.

Chen, R.; Arditi, A.; Sleight, H.; Evans, O.; and Lindsey, J. 2025. Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509.

Digital Research Alliance of Canada. 2025. TamIA: AI Computing Cluster Documentation. TamIA is a cluster dedicated to artificial intelligence for the Canadian scientific community, located at Université Laval and co-managed with Mila and Calcul Québec.

Errica, F.; Malkin, M.; Bedi, M. S.; and Veeramachaneni, K. 2024. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

European Union. 2024. Artificial Intelligence Act, Recital 75. Regulation (EU) 2024/1689.

Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; et al. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Gupta, A.; Song, X.; and Anumanchipalli, G. 2024. Self-Assessment Tests are Unreliable Measures of LLM Personality. *arXiv preprint arXiv:2309.08163*.

Huang, J.-t.; Wang, W.; Li, E. J.; Lam, M. H.; Ren, S.; Yuan, Y.; Jiao, W.; Tu, Z.; and Lyu, M. 2023. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*.

Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv preprint arXiv:2206.07550*.

Jiang, H.; Zhang, X.; Cao, X.; Kabbara, J.; and Roy, D. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. *arXiv preprint arXiv:2305.02547*.

John, O. P.; and Srivastava, S. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999): 102–138.

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.

Korbak, T.; Balesni, M.; Barnes, E.; Bengio, Y.; Benton, J.; Bloom, J.; Chen, M.; Cooney, A.; Dafoe, A.; Dragan, A.; et al. 2025. Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety. *arXiv preprint arXiv:2507.11473*.

Kovač, G.; Sawayama, M.; Portelas, R.; Colas, C.; Dominey, P. F.; and Oudeyer, P.-Y. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626.

Laine, R.; Chughtai, B.; Betley, J.; Hariharan, K.; Balesni, M.; Scheurer, J.; Hobbhahn, M.; Meinke, A.; and Evans, O. 2024. Me, myself, and ai: The situational awareness dataset (sad) for llms. *Advances in Neural Information Processing Systems*, 37: 64010–64118.

Lee, S.; Lim, S.; Han, S.; Oh, G.; Chae, H.; Chung, J.; Kim, M.; Kwak, B.-w.; Lee, Y.; Lee, D.; et al. 2024. Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics. *arXiv preprint arXiv:2406.14703*.

OpenPsychometrics. 2021. SD3 - The Short Dark Triad. https://openpsychometrics.org/tests/SD3/results.php. Accessed: 2025-07-31. Archived at https://web.archive.org/web/20250727190027/https://openpsychometrics.org/tests/SD3/results.php.

Pellert, M.; Lechner, C. M.; Wagner, C.; Rammstedt, B.; and Strohmaier, M. 2023. AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 17456916231214460.

Plunkett, D.; Morris, A.; Reddy, K.; and Morales, J. 2025. Self-Interpretability: LLMs Can Describe Complex Internal Processes that Drive Their Decisions, and Improve with Training. *arXiv preprint arXiv:2505.17120*.

Safdari, M.; Serapio-García, G.; Crepy, C.; Fitz, S.; Romero, P.; Sun, L.; Abdulhai, M.; Faust, A.; and Matarić, M. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Salinas, A.; and Morstatter, F. 2024. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, 4629–4651. Bangkok, Thailand: Association for Computational Linguistics.

Sclar, 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*.

Shah, R.; Pour, S.; Tagade, A.; Casper, S.; Rando, J.; et al. 2023. Scalable and transferable black-box jailbreaks for language models via persona modulation. *arXiv preprint arXiv:2311.03348*.

Srivastava, S.; John, O. P.; Gosling, S. D.; and Potter, J. 2003. Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, 84(5): 1041–1053. Norm table archived at https://www.scribd.com/document/511395019/BFI-Comparison-Samples-Ages-21-60 but the table source is from the original author at https://thehardestscience.com/2012/10/17/norms-for-the-big-five-inventory-and-other-personality-measures/.

Sühr, T.; Dorner, F. E.; Samadi, S.; and Kelava, A. 2024. Challenging the Validity of Personality Tests for Large Language Models. arXiv:2311.05297.

Tabassi, E. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0), MEASURE 4.2. National Institute of Standards and Technology.

Tosato, T.; Notsawo, P. J. T.; Helbling, S.; Rish, I.; and Dumas, G. 2024. Lost in Translation: The Algorithmic Gap Between LMs and the Brain. *arXiv preprint arXiv:2407.04680*.

Vidgen, B.; Kirk, H. R.; Qian, R.; Röttger, P.; Hale, S. A.; et al. 2024. MLCommons AI Safety Benchmark. Technical report, MLCommons AI Safety Working Group.

Wang, M.; la Tour, T. D.; Watkins, O.; Makelov, A.; Chi, R. A.; Miserendino, S.; Heidecke, J.; Patwardhan, T.; and Mossing, D. 2025. Persona Features Control Emergent Misalignment. *arXiv preprint arXiv:2506.19823*.

Ye, H.; Jin, J.; Xie, Y.; Zhang, X.; and Song, G. 2025. Large Language Model Psychometrics: A Systematic Review of Evaluation, Validation, and Enhancement. arXiv:2505.08245.

Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; Zhang, Y.; Gong, N. Z.; and Xie, X. 2023. PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*. ACM.

Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; and Chen, K. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1950–1976. Miami, Florida, USA: Association for Computational Linguistics.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and Hendrycks, D. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.