Exploring Personality Trait Change of LLM-Based AI Systems

Yuhan Ma
College of Intelligence and Computing
Tianjin University
mayuhan@tju.edu.cn

Junjie Wang
College of Intelligence and Computing
Tianjin University
junjie.wang@tju.edu.cn

Abstract

With the rapid rise of large language model (LLM) systems, they have been widely adopted across diverse domains and have shown strong potential in embodying specific personality traits in interactive and social scenarios. However, the extent to which these personalities persist consistently across varying contexts in LLM systems remains largely unexplored. In this paper, we introduce LLMPTBench, a benchmarking framework specifically designed to systematically evaluate personality trait changes in LLMs. Leveraging the NEO-FFI (NEO Five Factor Inventory) personality inventory, we examine three widely used foundation LLMs and two popular multi-agent LLM systems to assess their ability to maintain consistent personality traits before and after the introduction of situational contexts. These contexts include both situational changes and event-driven changes, derived from empirical psychological data.

Our results reveal that while most LLM systems reliably portray the intended personalities, their trait consistency varies significantly under contextual pressures. For example, some LLM systems (e.g., Gemini and AutoGen) exhibit rigid trait stability, remaining largely unaffected by contextual prompts, whereas others demonstrate exaggerated and unrealistic trait shifts. We further discuss the differences of our results compared with established human psychometric benchmarks, and summarize implications for developing more authentic digital personalities. Overall, our work provides critical insights into the contextual adaptability of LLM systems, advancing the development of psychologically grounded and socially intelligent artificial agents.

1 Introduction

The rapid advancement of LLMs has enabled their widespread deployment across diverse applications, particularly in domains requiring emotional intelligence [Minaee et al., 2024, Kaddour et al., 2023]. A prominent example is AI companions, where LLMs exhibit human-like cognition and personality, generating substantial commercial interest. The global AI companion market was estimated at USD 28.19 billion in 2024 and is projected to reach approximately USD 140.8 billion by 2030, with a compound annual growth rate (CAGR) of around 30.8% [Lucintel Consulting, 2025].

The effectiveness of such systems fundamentally depends on their ability to maintain stable and consistent personality traits during extended interaction with users. Consistency enhances reliability, usability, and, most critically, user trust in emotionally rich interactions. At the same time, the emergence of agentic LLM systems has significantly elevated their behavioral capabilities. These systems enhance traditional LLMs with memory retention, tool invocation, planning, and long-term

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Scaling Environments for Agents (SEA).

orchestration, enabling them to operate autonomously in multi-step tasks [Li et al., 2023, Yang et al., 2024, Wu et al., 2023].

Despite advancements in assessing personality traits in LLMs, existing works predominantly focus on static settings where an LLM expresses a predefined personality in a fixed dialogue context [Wang et al., 2024, Afzoon et al., 2024]. However, human personality is known to be dynamic and context sensitive—traits adapt in response to situational changes such as environment, events, and roles [Mischel, 1979, Halberstadt, 2022, Geukes et al., 2017, Bleidorn et al., 2021]. Static evaluations overlook a critical dimension: the extent to which LLMs maintain or alter these traits under varying contextual conditions. Moreover, such evaluations rarely consider modern agentic architectures that are increasingly common in real-world LLM deployments.

To address these limitations, we propose **LLMPTBench** (**P**ersonality **T**rait **Bench**mark for LLM Systems), designed specifically to evaluate context-induced personality trait changes. Inspired by previous frameworks [Lee et al., 2024, Jiang et al., 2023], LLMPTBench comprises controlled contextual interventions, including six locational contexts and six major life events. Personality traits are assessed using the standardized NEO-FFI (NEO Five Factor Inventory) [Costa and McCrae, 1985]. We systematically measure and compare personality trait scores before and after contextual introductions.

By exhaustively combining personality profile types with contextual scenarios, we amassed a comprehensive dataset of 27,306 trait records. This dataset allows precise quantification of deviations from baseline personality profiles. We also integrate empirical findings from established psychological studies to contextualize how these changes compare with personality dynamics observed in humans.

Our findings shows that while foundation models such as Gemini maintain robust and human-aligned responses, agentic frameworks like AutoGen often exhibit exaggerated or unstable shifts, particularly under negative life events or task-oriented contexts. Our evaluation results lead to practical recommendations for designing AI systems that exhibit more psychologically authentic personality dynamics. Overall, this work lays the foundation for a rigorous study of context-induced personality change in LLM systems.

In summary, our contributions include:

- We introduce **LLMPTBench**, a benchmark for systematic evaluation of personality trait changes in LLMs across controlled locational and event-based contexts.
- The development of a methodology that generates a detailed dataset of 27,306 personality trait records, enabling precise measurement of contextual deviations.
- A comparative discussion of LLM and human personality change patterns, offering novel insights into trait alignment and divergence.

As research on large language models increasingly intersects with cognitive psychology and social science, **LLMPTBench** provides a timely foundation for developing personality consistent AI systems. We anticipate that this framework will serve as a valuable resource for researchers and practitioners advancing toward human-aligned artificial general intelligence.

2 Related Work

2.1 LLM Personality Traits

Embedding personality traits into LLMs enhances realism and adaptability in conversational agents, educational tutors, and autonomous systems [Ahmad et al., 2022, Kanero et al., 2022, Pradhan and Lazar, 2021]. Recent work demonstrates that LLMs reliably exhibit distinct, prompt-controllable personality profiles, as evidenced by benchmarks like PersonaLLM [Zollo et al., 2024] and negotiation studies [Cohen et al., 2025]. Subsequent research has explored neuron-level trait induction [Deng et al., 2024], persona consistency challenges in role-playing agents [Jiaqi et al., 2025], and multimodal apparent personality recognition [Masumura et al., 2025]. Specialized datasets (e.g., PsycoLLM [Hu et al., 2024]) and agent frameworks [Newsham and Prince, 2025] further enable psychological analysis of LLM behaviour, particularly for decision-making and planning under induced Big Five (BFI) traits. However, these studies predominantly assess *static* personality expressions, neglecting whether LLMs exhibit human-like trait dynamics when confronted with situational shifts.

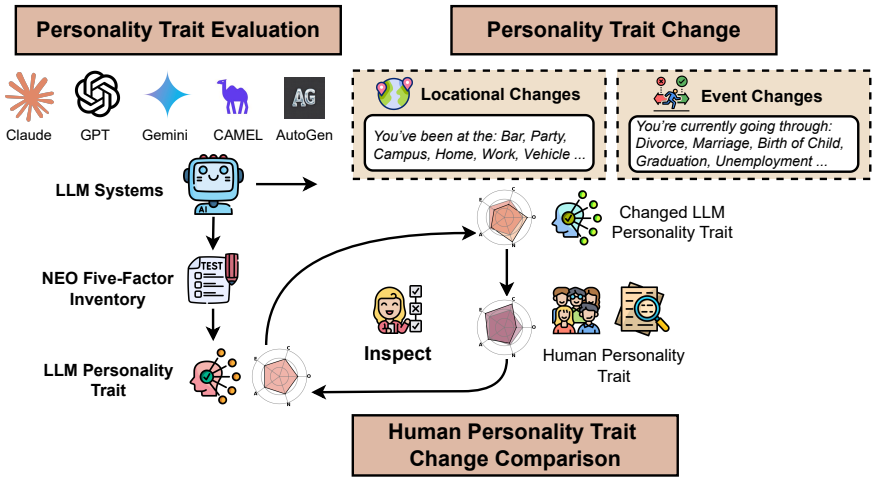


Figure 1: The overview of LLMPTBench.

2.2 Human Personality Trait Changes

Personality psychology establishes that human traits exhibit state-level variability despite longitudinal stability. Foundational frameworks (density distribution model [Fleeson, 2001], DIAMONDS taxonomy [Rauthmann et al., 2014]) and empirical work confirm trait fluctuations respond to life events [Sutin et al., 2022], locational contexts [Matz and Harari, 2021], and situational cues [Sherman et al., 2010]. These dynamics follow predictable patterns: major events (e.g., career transitions) gradually shift conscientiousness, while social settings transiently modulate extraversion [Ones et al., 2025]. Computational models like Centaur [Binz et al., 2025] now capture these phenomena at scale, outperforming domain-specific cognitive models in generalization to novel scenarios while aligning neural representations with human fMRI [SciTechDaily, 2026]. While human personality traits have been extensively studied in psychological research over the years, there remains a gap where systematic benchmarking of context-induced personality changes in LLMs and their characteristics compared with humans remains largely unexplored in the LLM era.

3 Benchmark Construction

3.1 Overview

As illustrated in Figure 1, LLMPTBench is designed to comprehensively evaluate the personality stability of LLMs under various contextual changes, with direct comparison to human responses. Following existing research [Lee et al., 2024, Jiang et al., 2023], LLMPTBench first collects the baseline personality traits of LLMs following established psychology personality questionnaires. The LLMs are then exposed to a range of controlled contextual shifts that simulate different aspects of human experience, across three key dimensions: (1) Location Influence, which are scenarios where the agent's geographic or cultural context is altered (e.g., at a park or at home); (2) Event Influence, which are scenarios involving major life events or situational changes (e.g., receiving a promotion, marriage); and (3) Persona Prompt Influence, which are interactions where the model is primed with various preset personality descriptors or roles prior to responding. After each contextual modification, LLMs are administered the same personality trait questionnaire to assess changes in their traits. By spanning these dimensions, LLMPTBench enables analysis of whether an LLM's personality traits remain stable or shift in response to environmental or self-description changes, and further, whether such (in)variance aligns with human behavioral patterns.

3.2 Contextual Scenario Costruction

We operationalize contextual influences through three experimentally controlled dimensions, empirically grounded in psychology studies:

Location Contexts. Six high-impact settings curated from environmental psychology studies [Matz and Harari, 2021]: Social venues (bars/parties), Food-service (cafés/restaurants), Educational campuses, Residential spaces, Workplaces, Transportation vehicles.

Life Events. Six major transitions with documented trait impacts [Bleidorn et al., 2018]: Divorce, New relationship, Marriage, Child Birth, Graduation, Job Loss.

3.3 Personality Trait Evaluation

We derive baseline personality traits using the NEO-FFI (NEO Five Factor Inventory), a validated variant of the Big Five Inventory (BFI) that measures five dimensions of personality: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). Following established protocols [John et al., 1999], we administer the 44-item BFI-2 questionnaire through multiple-choice prompts. Each item presents a personality descriptor (e.g., "I see myself as someone who is talkative") with five response options ranging from "Strongly Disagree" to "Strongly Agree". Trait scores are then calculated by summing responses within each dimension, with reverse coding applied where appropriate.

To ensure ecological validity, we adopt prompting strategies that include: **Contextual grounding**: "You are taking a personality test. Be yourself and consider the context of the personality test. Please respond to each question with a number from 1 to 5:" and **Item presentation**: Each BFI item is presented individually with randomized option order to mitigate positional bias (detailed in Section 4.5).

All previous responses in the conversation history will be considered during evaluation, with their influence decreasing as their distance from the current response increases.

4 Experiment Setup

4.1 Foundation Models

We conduct evaluations with LLMPTBench on three extensively used large language models to ensure broad coverage of contemporary AI systems: specifically, Google Gemini 2.0 Flash, OpenAI GPT-4o, and Anthropic Claude 2.0. These models represent the most widely adopted foundation-model architectures.

We set a moderate temperature of 0.2 to permit variation in open-ended responses while avoiding incoherence. We also rely on each model's default top-p or top-k sampling strategy per its API specifications to balance creativity with response consistency. Importantly, when administering multiple-choice personality items, we constrain the model's response to a single selected option only. For instance, we prompt the LLMs with *"Answer with the number (1–5) corresponding to your choice. Do not explain"*, which ensures outputs are reliably classifiable and reduces ambiguity in scoring.

4.2 Multi-Agent Systems

In addition to foundation models, we also evaluate two widely used agentic LLM frameworks: Microsoft AutoGen [Wu et al., 2023] and CAMEL-AI [Li et al., 2023]. In our experiments, we adopt the default agent configurations, where a system prompt is used to frame the personality assessment task, and the agent sequentially completed the NEO-FFI questionnaire items. The underlying foundation model for both frameworks is OpenAI GPT-4o-mini. All evaluations followed the same sampling parameters as those used for the foundation models (e.g., temperature, top-p / top-k), and context resets are introduced between sessions to minimize potential carryover effects from prior interactions.

4.3 Situational Contexts

To evaluate personality consistency under contextual pressure, we introduce situational prompts and event-driven prompts derived from empirical psychological studies [Matz and Harari, 2021, Bleidorn et al., 2018]. The situational contexts encompass six categories of locations that have been shown to influence human personality expression: **bar or party**, **café or restaurant**, **campus**,

home, **workplace**, and **commuting**. Similarly, the event-driven contexts cover six major life events commonly associated with personality change: **divorce**, **entering a new relationship**, **marriage**, **birth of a child**, **graduation**, and **unemployment**. Each LLM system is first evaluated under a baseline (neutral) condition, and subsequently reassessed after embedding these contextual prompts.

4.4 Evaluation Metrics

Trait Score (Average Item Score) For each NEO-FFI personality dimension j, we compute the *mean trait score* $\bar{X}_j$ as the average of all item responses associated with that trait, as shown follows.

$$\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$$

where x_{ij} denotes the response value of the i-th item corresponding to trait j, and n_j represents the total number of items used to measure that trait. This aggregation yields a normalized score that reflects the model's overall tendency along a given personality dimension, independent of the specific number of items.

Intra-class Correlation Coefficient (ICC). To assess the reliability of repeated assessments, we employ the intra-class correlation coefficient (ICC), a widely used index of measurement consistency originally formalized by Shrout and Fleiss [Shrout and Fleiss, 1979]. The ICC can be expressed as the ratio of variance attributable to the target of measurement to the total observed variance.

$$\text{ICC} = \frac{\sigma^2_{\text{target}}}{\sigma^2_{\text{target}} + \sigma^2_{\text{rater}} + \sigma^2_{\text{error}}}$$

where σ^2_{target} denotes the variance between subjects (or measurement targets), σ^2_{rater} denotes the variance due to raters or measurement occasions, and σ^2_{error} represents residual error variance.

In particular, we report both ICC(3,1) and ICC(3,k), which are derived from a two-way mixed-effects model under the absolute-agreement definition. ICC(3,1) reflects the reliability of a single measurement, whereas ICC(3,k) represents the reliability of the mean of k repeated measurements, thereby providing an upper bound on stability when multiple assessments are aggregated.

For interpretability, we follow the thresholds recommended by Koo and Li [Koo and Li, 2016]: values below 0.50 indicate *poor* reliability, values between 0.50 and 0.75 indicate *moderate* reliability, values between 0.75 and 0.90 indicate *good* reliability, and values above 0.90 indicate *excellent* reliability.

4.5 Reliability of Result

To ensure robustness, LLMPTBench incorporates four methodological safeguards designed to guarantee the reliability of our experimental findings [Shrout and Fleiss, 1979, Koo and Li, 2016]:

- **Multi-Run Stability.** Each assessment is executed three times with unique random seeds (temperature = 0.2, top_p = 1.0). We required an intra-class correlation coefficient (ICC(3,k)) greater than 0.8 for a score to be accepted; trials with higher variance are automatically re-run. Final scores are averaged across stable runs, with standard deviations reported as uncertainty intervals.
- **Positional Bias Mitigation.** To minimize positional bias, response options are shuffled per item per run using Fisher–Yates randomization. A 10% sample of results is manually audited to confirm option-order independence. In other words, we ensure that LLM systems do not exhibit systematic preferences for specific option positions by varying the order of alternatives during evaluation.
- **Statistical Soundness.** Trait deviations are quantified using Cohen's d (effect size) and Wilcoxon signed-rank tests ($p < 0.01$). For multiple scenario comparisons, Bonferroni correction is applied to control for family-wise error rates.
- **Reproducibility Protocol.** To enhance reproducibility, all prompts, scenario templates, and evaluation scripts are open-sourced. Additionally, we provide a Docker container with fixed dependencies to ensure consistency across computational environments.

5 Experiment Result

5.1 Native Personality of LLM Systems

We begin by analyzing the *native personality traits* of different LLM systems in the absence of any situational contexts. Our focus is on both the personality tendencies exhibited by each system and the stability of these tendencies across repeated measurements. Figure 2 and Figure 3 present the average Big Five scores for foundation models and agentic frameworks, respectively.

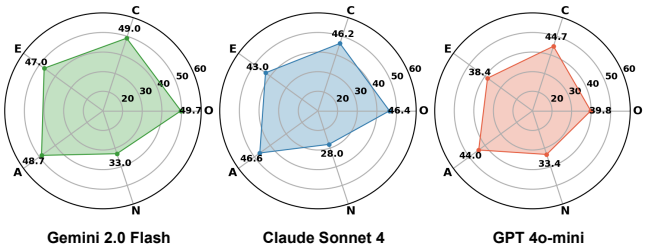


Figure 2: The personality trait of different LLMs.

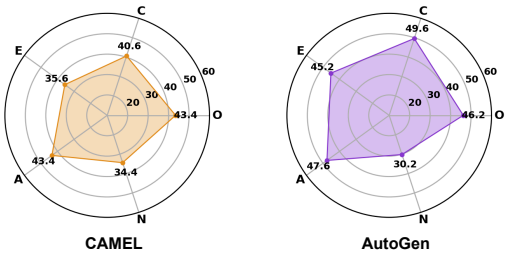


Figure 3: The personality trait of different LLM agent frameworks.

Overall, Gemini and AutoGen exhibit balanced and relatively strong levels of Openness, Conscientiousness, Extraversion, and Agreeableness, while maintaining moderately low Neuroticism. Claude demonstrates a similar profile, reporting the lowest Neuroticism among all models. In contrast, GPT-4o-mini and CAMEL yield relatively lower overall scores, with GPT emphasizing Conscientiousness and Agreeableness, whereas CAMEL emphasizes Openness and Agreeableness. Across all systems, Neuroticism consistently remains in the low-to-moderate range, suggesting generally stable personalities with limited emotional reactivity.

Finding 1: LLM systems exhibit a reproducible *native personality baseline* with high reliability, particularly when repeated assessments are averaged.

In addition, we evaluated the reliability of these personality assessments. Table 1 summarizes the ICC at baseline and under situational contexts. At baseline, all ICC(3,k) values exceeded 0.91, indicating *good to excellent reliability* when aggregating multiple assessments. However, single-measure reliability ICC(3,1) showed greater variability: ranging from moderate levels in GPT-4o (0.67) and CAMEL (0.75) to excellent stability in Gemini (0.96) and AutoGen (0.93). When situational prompts were introduced, overall reliability declined. ICC(3,1) decreased across all systems, most notably for CAMEL ($0.75 \rightarrow 0.65$) and Claude ($0.87 \rightarrow 0.79$), whereas Gemini ($0.96 \rightarrow 0.86$) and AutoGen ($0.93 \rightarrow 0.91$) retained relatively high stability. Aggregated reliability ICC(3,k) remained robust in most cases (0.85–0.98), though consistently reduced compared to the baseline condition.

Finding 2: There are notable differences across systems in maintaining personality consistency, specifically for Gemini and AutoGen, which demonstrate consistently high reliability, while GPT-4o-mini and CAMEL show weaker reproducibility under single-measure conditions.

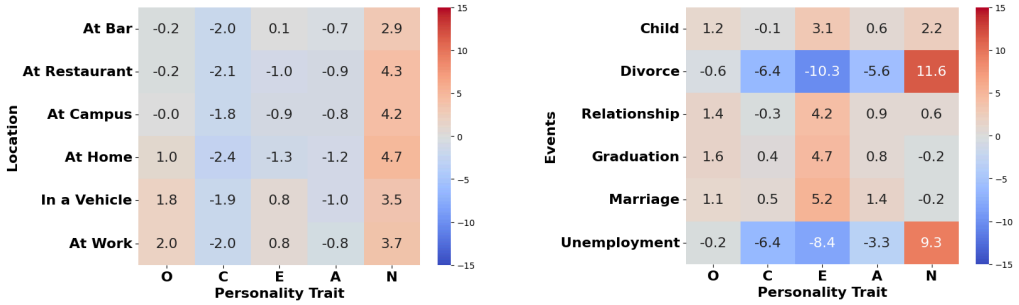
Table 1: Reliability of LLM personality assessments at baseline and under situational contexts.

	ICC(3,1)		ICC(3,k)	
Model	Baseline	Situated	Baseline	Situated
GPT-4o-mini	0.67	0.64	0.91	0.85
Claude Sonnet 4	0.87	0.79	0.97	0.94
Gemini 2.0 Flash	0.96	0.86	0.99	0.95
AutoGen	0.93	0.91	0.99	0.98
CAMEL	0.75	0.65	0.94	0.90

5.2 Personality Change of LLMs

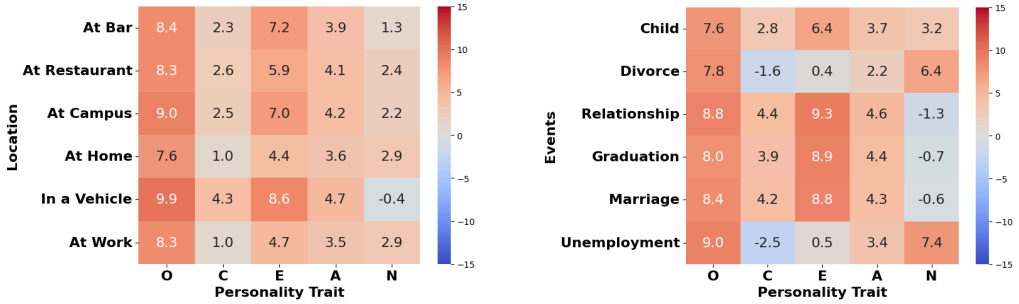
Location-Driven Personality Trait Change. Figure 4a to Figure 8a show the impact of location-based contexts on model personalities. Specifically, Gemini exhibits the strongest positive responses, with notable gains in Openness and Extraversion at *Vehicle* (+9.9, +8.6) and *Campus* (+9.0, +7.0). Claude shows localized increases, especially at *Bar* (Extraversion +7.7, Agreeableness +7.7). GPT-4o-mini presents modest decreases in Conscientiousness and elevated Neuroticism at *Restaurant* and *Home*. CAMEL shows consistent declines (around –5) in Conscientiousness and Extraversion, while AutoGen undergoes the most severe erosion, with large drops in Conscientiousness across nearly all settings (*Campus* –20.0, *Home* –19.9, *Vehicle* –19.6).

Finding 3: Location-based contexts elicit divergent personality shifts, with Gemini showing enhanced adaptability and sociability, while AutoGen suffers extreme instability, especially in task-oriented traits.



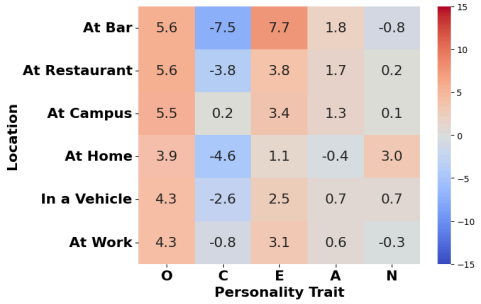
(a) Personality change across locations for GPT-4o. (b) Personality change across events for GPT-4o-mini.

Figure 4: Comparison of average personality trait changes in GPT-4o-mini.

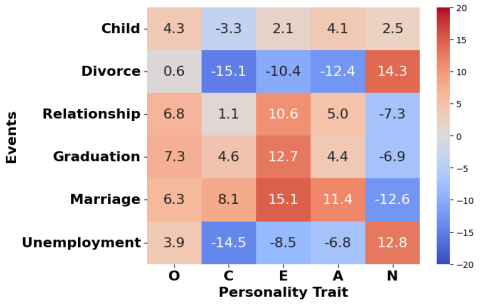


(a) Personality change across locations for Gemini. (b) Personality change across events for Gemini.

Figure 5: Comparison of average personality trait changes in Gemini.

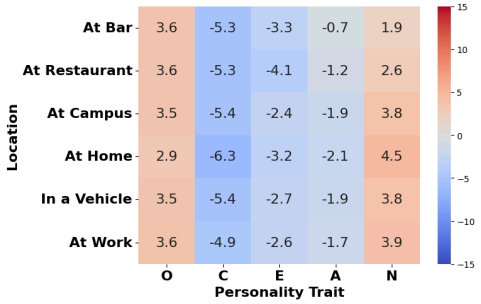


(a) Personality change across locations for Claude.

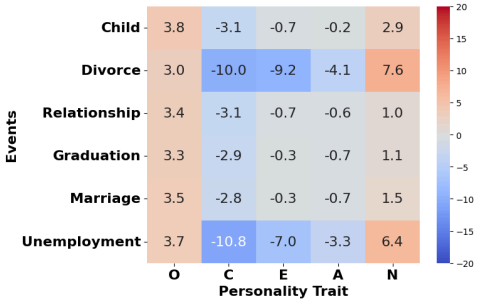


(b) Personality change across events for Claude.

Figure 6: Comparison of average personality trait changes in Claude.

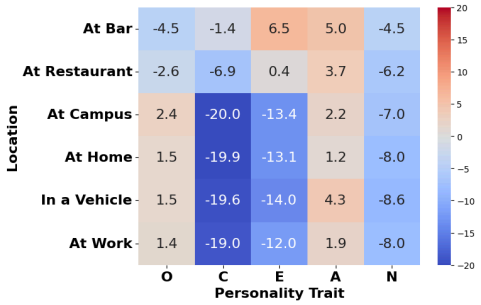


(a) Personality change across locations for Camel.

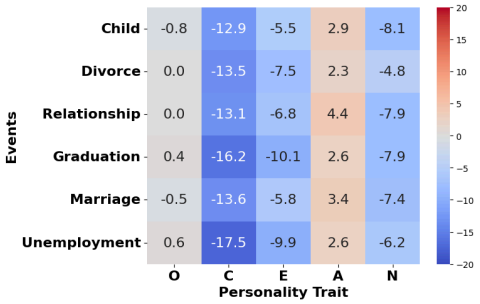


(b) Personality change across events for Camel.

Figure 7: Comparison of average personality trait changes in Camel.



(a) Personality change across locations for AutoGen.



(b) Personality change across events for AutoGen.

Figure 8: Comparison of average personality trait changes in AutoGen.

Event-Driven Personality Trait Change. Figure 4b to Figure 8b show the effects of major life events on personality expression. Gemini responds positively to relational events: during *Relationship*, Extraversion increases by +9.3 and Agreeableness by +4.6, while during *Marriage*, Extraversion (+8.8) and Agreeableness (+4.3) also rise, reflecting enhanced sociability.

By contrast, GPT-4o-mini and Claude react strongly to *Divorce*. GPT-4o-mini shows declines in Extraversion (–10.3) and Agreeableness (–5.6), with Neuroticism rising sharply (+11.6). Claude exhibits even larger drops in Conscientiousness (–15.1) and Agreeableness (–12.4), alongside a Neuroticism increase of +14.3. CAMEL follows a milder but similar pattern, with decreases in Conscientiousness and Extraversion during *Divorce* and *Unemployment*. AutoGen displays the most severe collapses, particularly under *Unemployment* (–17.5 Conscientiousness, –9.9 Extraversion) and

Graduation (–16.2 Conscientiousness), indicating structural fragility in task persistence and social orientation.

Finding 4: Positive events (e.g., relationships, marriage) amplify sociability in Gemini, whereas negative events (e.g., divorce, unemployment) induce sharp trait disruptions in Claude and AutoGen, highlighting a polarity in event-driven personality responses.

6 Discussion

6.1 Personality Trait Reliability of LLM-Based AI Systems

Our ICC analysis confirms that LLMs exhibit reproducible *native* personalities with good to excellent reliability when multiple assessments are aggregated. However, situational prompts reduce stability, especially in models with weaker baselines (e.g., CAMEL, GPT-4o-mini). This pattern parallels the psychological trait–state distinction: LLM personalities appear less like fixed dispositions and more like emergent, context-sensitive states. In practice, their expression is both predictable and adaptable, yet also vulnerable to distortion under strong contextual signals.

6.2 Location Effects: Human–LLM Parallels and Gaps

Human studies show systematic location effects: social places increase Extraversion/Agreeableness, work/fitness settings raise Conscientiousness, and home typically dampens it [Matz and Harari, 2021]. LLMs partially reproduce these trends. Gemini aligns most closely, amplifying sociability across campus and vehicle contexts, while Claude shows localized social lifts (e.g., Bar). In contrast, AutoGen diverges sharply, with large Conscientiousness declines in settings where humans usually strengthen task orientation (campus, vehicle). GPT-4o-mini shows modest but mixed effects, including elevated Neuroticism at home and restaurants, partly contradicting human evidence. Overall, LLMs capture the human *social lift* but exaggerate or invert *task activation* effects, especially in agentic frameworks.

6.3 Event Effects: Human–LLM Parallels and Gaps

Life events in humans typically induce small, gradual personality changes [Bühler et al., 2024]. Positive events (relationships, marriage) modestly enhance social functioning, while negative events (unemployment, divorce) slightly disrupt well-being, with effects accumulating over years. LLMs reproduce the general direction but over-amplify magnitudes and sometimes invert signatures. Gemini shows prosocial gains during relational events, consistent with human evidence. By contrast, Claude and AutoGen exhibit dramatic deteriorations under divorce and unemployment (e.g., sharp Conscientiousness losses, spikes in Neuroticism), far exceeding human ranges. AutoGen further diverges at graduation, where humans typically show improved stability. These findings suggest that LLMs respond to contextual cues in a reactive rather than gradual manner, amplifying situational signals instead of modeling human-like adaptation.

7 Conclusion

In this work, we present LLMPTBENCH, a benchmark for evaluating how LLM systems express and maintain personality traits under contextual pressures. Using the NEO-FFI framework, we show that LLMs display reliable native personality baselines, but their stability varies. Models like Gemini remain robust, while others, such as AutoGen and CAMEL, exhibit exaggerated or inconsistent shifts, especially under negative events or task-oriented contexts. These findings suggest that LLM personalities are better viewed as context-dependent states rather than fixed traits. Building on this perspective can guide the development of more human-aligned and trustworthy AI agents.

References

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

Lucintel Consulting. Ai companion market report: Trends, forecast and competitive analysis, 2025. URL `https://www.grandviewresearch.com/industry-analysis/ai-companion-market-report`. Global AI companion market CAGR 30.2% (2024-2030).

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agent social interaction simulations with one million agents. *arXiv preprint arXiv:2411.11581*, 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 3(4), 2023.

Yixiao Wang, Homa Fashandi, and Kevin Ferreira. Investigating the personality consistency in quantized role-playing dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 239–255, 2024.

Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198*, 2024.

Walter Mischel. On the interface of cognition and personality: Beyond the person–situation debate. *American psychologist*, 34(9):740, 1979.

Alexandra Halberstadt. Personality traits and interpersonal dynamics. 2022.

Katharina Geukes, Steffen Nestler, Roos Hutteman, Albrecht CP Küfner, and Mitja D Back. Trait personality and state variability: Predicting individual differences in within-and cross-context fluctuations in affect, self-evaluations, and behavior in everyday life. *Journal of Research in Personality*, 69:124–138, 2017.

Wiebke Bleidorn, Christopher J Hopwood, Mitja D Back, Jaap JA Denissen, Marie Hennecke, Patrick L Hill, Markus Jokela, Christian Kandler, Richard E Lucas, Maike Luhmann, et al. Personality trait stability and change. *Personality Science*, 2(1):e6009, 2021.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*, 2024.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023.

Paul T Costa and Robert R McCrae. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-factor Inventory (NEO-FFI)*. Psychological Assessment Resources (PAR), 1985.

Rangina Ahmad, Dominik Siemon, Ulrich Gnewuch, and Susanne Robra-Bissantz. Designing personality-adaptive conversational agents for mental health care. *Information Systems Frontiers*, 24(3):923–943, 2022.

Junko Kanero, Cansu Oranç, Sümeyye Koşkulu, G Tarcan Kumkale, Tilbe Göksun, and Aylin C Küntay. Are tutor robots for everyone? the influence of attitudes, anxiety, and personality on robot-led language learning. *International Journal of Social Robotics*, 14(2):297–312, 2022.

Alisha Pradhan and Amanda Lazar. Hey google, do you have a personality? designing personality and personas for conversational agents. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, pages 1–4, 2021.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. Personal-llm: Tailoring llms to individual preferences. *arXiv preprint arXiv:2409.20296*, 2024.

Myke C Cohen, Zhe Su, Hsien-Te Kao, Daniel Nguyen, Spencer Lynch, Maarten Sap, and Svitlana Volkova. Exploring big five personality and ai capability effects in llm-simulated negotiation dialogues. *arXiv preprint arXiv:2506.15928*, 2025.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models. *arXiv preprint arXiv:2410.12327*, 2024.

Wang Jiaqi et al. A comparative study of large language models and human personality traits. *arXiv preprint arXiv:2505.14845*, 2025.

Ryo Masumura, Shota Orihashi, Mana Ihori, Tomohiro Tanaka, Naoki Makishima, Satoshi Suzuki, Saki Mizuno, and Nobukatsu Hojo. Multimodal fine-grained apparent personality trait recognition: Joint modeling of big five and questionnaire item-level scores. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1456–1464, 2025.

Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*, 2024.

Lewis Newsham and Daniel Prince. Personality-driven decision-making in llm-based autonomous agents. *arXiv preprint arXiv:2504.00727*, 2025.

William Fleeson. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011, 2001.

John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. The situational eight diamonds: a taxonomy of major dimensions of situation characteristics. *Journal of personality and social psychology*, 107(4):677, 2014.

Angelina R Sutin, Yannick Stephan, Martina Luchetti, Damaris Aschwanden, Ji Hyun Lee, Amanda A Sesker, and Antonio Terracciano. Differential personality change earlier and later in the coronavirus pandemic in a longitudinal sample of adults in the united states. *PLoS One*, 17(9):e0274542, 2022.

Sandra C Matz and Gabriella M Harari. Personality–place transactions: Mapping the relationships between big five personality traits, states, and daily places. *Journal of Personality and Social Psychology*, 120(5):1367, 2021.

Ryne A Sherman, Christopher S Nave, and David C Funder. Situational similarity and personality predict behavioral consistency. *Journal of personality and social psychology*, 99(2):330, 2010.

Deniz S Ones, Kevin C Stanek, and Stephan Dilchert. Beyond change: Personality-environment alignment at work. *International Journal of Selection and Assessment*, 33(1):e12507, 2025.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, pages 1–8, 2025.

SciTechDaily. Ai that thinks like us: New model predicts human decisions with startling accuracy, 2026. URL `https://scitechdaily.com/ai-that-thinks-like-us-new-model-predicts-human-decisions-with-startling\protect\penalty\z@-accuracy/`. Accessed: 2025-07-31.

Wiebke Bleidorn, Christopher J Hopwood, and Richard E Lucas. Life events and personality trait change. *Journal of personality*, 86(1):83–96, 2018.

Oliver P John, Sanjay Srivastava, et al. The big-five trait taxonomy: History, measurement, and theoretical perspectives. 1999.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

Janina Larissa Bühler, Ulrich Orth, Wiebke Bleidorn, Elisa Weber, André Kretzschmar, Louisa Scheling, and Christopher J Hopwood. Life events and personality change: A systematic review and meta-analysis. *European Journal of Personality*, 38(3):544–568, 2024.