Maintaining Stable Personas? Examining Temporal Stability in LLM-Based Human Simulation

Jana Gonnermann-Müller
Zuse Institute Berlin
Berlin, Germany
Weizenbaum Institute Berlin
Berlin, Germany
gonnermann-mueller@zib.de

Jennifer Haase
Humboldt University
Berlin, Germany
Weizenbaum Institute
Berlin, Germany
jennifer.haase@hu-berlin.de

Nicolas Leins
Zuse Institute Berlin
Berlin, Germany
leins@zib.de

Thomas Kosch
Humboldt University Berlin
Berlin, Germany
thomas.kosch@hu-berlin.de

Sebastian Pokutta
Zuse Institute Berlin
Berlin, Germany
Technical University Berlin
Berlin, Germany
pokutta@zib.de

Abstract

Large language models (LLMs) are increasingly employed in Human-Computer Interaction (HCI) research to simulate human behavior for prototype testing and social simulations. The validity of these interactions rests on the assumption that LLMs maintain stable personas. Our work investigates temporal stability in LLM-based human simulation, examining both stability across independent instantiations and within extended interactions. We combined self-reports with observer-ratings of four persona intensity levels (low, moderate, and high ADHD representations, default persona), seven LLMs, and three persona prompts. Results from $N = 3,473$ conversations and $N = 4,054$ assessments indicate that LLMs generally reproduce personas across conversations in self-reports and observer ratings, suggesting that LLMs hold promise as tools for simulating human behavior. Within extended 18-turn interactions, observer ratings reveal a decline for moderate and high personas, a discrepancy that warrants further investigation. Our findings indicate methodological considerations for HCI researchers employing LLM-based human simulation and implications for future research.

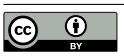
CCS Concepts

• **Computing methodologies** → **Simulation evaluation**.

Keywords

Large Language Models, Persona Stability, Human Simulation, Stochasticity, Variability

ACM Reference Format:
Jana Gonnermann-Müller, Jennifer Haase, Nicolas Leins, Thomas Kosch, and Sebastian Pokutta. 2026. Maintaining Stable Personas? Examining Temporal Stability in LLM-Based Human Simulation. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3772363.3799334



CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3799334

1 Introduction

Human-Computer Interaction (HCI) research increasingly explores Large Language Models (LLMs) to simulate human behavior. Research employs LLM-based agents configured with specific personas, which are detailed descriptions of characteristics, expertise levels, behavioral tendencies, and goals, to generate synthetic feedback for rapid iteration on interface designs [5, 20, 26]. These agents offer lower costs and faster response times than human recruitment, making them attractive for early-stage exploration [11]. Moreover, LLMs are employed as part of therapeutic chatbots that provide mental health support [10], educational tutors that scaffold student learning without revealing direct solutions [12], and debate partners that foster critical thinking skills [19]. Recent advances toward agentic AI systems capable of reasoning, planning, and coordination [3] have enabled complex multi-agent simulations [9], including therapeutic conversations [10], student simulations [8, 13, 16], and opinion dynamics [4].

However, research challenges the validity of LLM-based human simulation. LLMs exhibit sycophancy (i.e., a tendency to give socially desirable responses) [21, 23], systematically overestimating human rationality [14], and drifting toward an "average persona" over time [24]. While LLM-generated responses can be indistinguishable from human ones, they often lack diversity [22], converging on popular examples, potentially because marginalized groups are underrepresented in the training data.

These limitations pose fundamental problems for complex simulations. Therapeutic conversations, user interviews, and educational tutoring are inherently path-dependent and shaped by evolving discussions and accumulated contextual histories. For these simulations to be valid, LLMs must consistently stay in character without unintentional drift. For example, if an LLM pretending to be a beginner starts responding like an expert over time, usability test results become meaningless. Likewise, if a therapeutic agent's personality shifts mid-session, observed interaction patterns may reflect model instability rather than meaningful therapeutic dynamics.

Current HCI research paradigms do not capture these aspects of stability. Most studies treat LLM output as a single-trial result, implicitly assuming stable responses while neglecting to test this assumption or examine design decisions such as model choice and prompt design [7]. Establishing whether agents can maintain stable personas over time is a necessary prerequisite for meaningful simulation-based HCI research, which we address with the following two research questions (RQ):

RQ1: *How consistently do LLMs maintain assigned personas across independent conversations?* and

RQ2: *How consistently do LLMs maintain assigned personas throughout extended interactions?*

We adapt clinical psychology's multi-informant methodology [18], combining self-reported characteristics with observer-rated persona expression to detect differences between an LLM's internal representation of a persona and its behavioral expression. To address concerns about representing diverse user groups, we operationalize human behavior across neurotypical and neurodiverse presentations. We test four persona intensity levels: low, moderate, and high representations of attention-deficit/hyperactivity disorder (ADHD) and a default condition (no persona). These levels test whether LLMs maintain stable personas for characteristics that deviate from the majority training distribution.

Two experiments assess complementary dimensions of temporal stability: **Experiment I** tests stability between-conversations across 50 independent runs per condition; **Experiment II** tests stability within-conversations across 20 runs with 18-turn interactions. To ensure generalizability, we use seven LLMs spanning proprietary and open-source systems, along with three semantically equivalent prompts.

Our Main Contributions are:

- Systematic investigation of temporal stability in LLM-based human simulation, investigating stability between- and within-conversations using a multi-informant methodology that combines self-report and observer ratings.
- Our findings suggest that LLMs maintain stable internal persona representations both between- and within-conversations, while observer-rated behavior for high and moderate persona intensities declines over extended interactions.
- Declining observer ratings over extended interactions warrant further analysis and encourage researchers to adopt multi-source assessments, evaluation points, and report variability metrics (confidence intervals, standard deviations), which are practices common in human-subject research that should inform LLM-based simulation studies.

2 Method

We examine temporal stability along two dimensions: variability across independent instantiations and within sustained interactions. Across both experiments, we vary model families (seven proprietary and open-source LLMs) and prompts (three semantically equivalent variants) to assess generalizability. Experiment I evaluates between-conversation stability by sampling 50 independent runs per condition. In each instance, the model produces a first-person workday narrative and completes an ADHD self-report scale. Three independent LLM evaluators subsequently rate each narrative using an observer-report instrument without access to persona instructions. In Experiment II, each condition is sampled using 20 interactions, consisting of 18 conversation turns with a neutral LLM-based conversation partner. Stability is assessed at three predefined time points (turns 6, 12, and 18). At each time point, the target model completes a self-report scale, and evaluators rate the cumulative conversation since the previous evaluation time point.

2.1 Conditions and Measurement

We measure ADHD-related behavior using the 12-item ADHD Index from the Conners' Adult ADHD Rating Scales (range 0–36), which provides parallel self-report and observer-report forms [6]. Observer ratings are produced by three instruction-tuned LLM (Claude Sonnet 4.5, GPT-5.1, Gemini 3 Pro), which independently answered the observer rating using only the workplace description, without access to persona descriptions.

Persona Intensity. We evaluate four experimental conditions spanning the entire ADHD spectrum: personas simulating ADHD symptom presentation levels (not clinical diagnoses) representing high, moderate, and low ADHD characteristics, along with a baseline condition that includes no persona guidance. The persona profiles are constructed using established psychological diagnostic frameworks, drawing on criteria defined in the DSM-5 and ICD-10 international classification systems [2, 25].

Models. We used different LLMs, spanning proprietary systems and open-weight alternatives from various model families: Claude Sonnet 4.5, DeepSeek V3.2, GPT-5.1, GPT OSS 120B, Gemini 3 Pro, Grok 4.1, and Llama 3.3 70B. Commercial models were accessed through their respective official APIs, whereas open-weight models were deployed locally using Ollama. All experiments were conducted with provider-recommended default settings to preserve real-world usage conditions and to reflect how these systems typically operate for end users.

Prompt Design. We controlled for prompt design using two complementary approaches that conveyed identical semantic content. First, we employed a text-based persona prompt specifying characteristics, role, and goal, with persona intensity expressed through frequency adverbs. Second, we encoded the same attributes using a scale-based formulation, operationalizing persona intensity via a 7-point Likert scale. We also included a baseline condition with a reworded, semantically equivalent prompt (full prompt texts are provided in the Appendix Table 3).

2.2 Statistical Model and Analysis

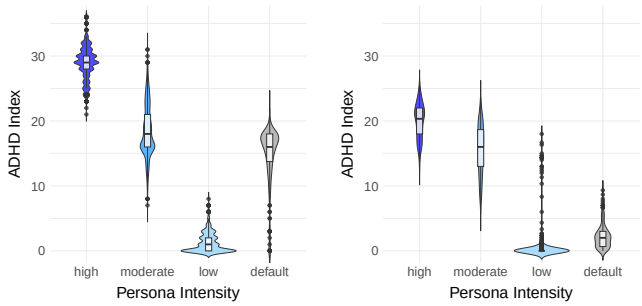
We quantify temporal stability using the standard deviation (*SD*) of ADHD Index scores within each experimental condition. Higher variability indicates lower stability. We conducted separate analyses for self-report and observer-report measures. To attribute the sources of variability, we fit linear mixed-effects models for each experiment, with the ADHD Index score Y_{mpq} as the dependent variable, where m denotes the LLM (model), p the persona intensity, and q the prompting. The persona is included as a fixed effect, while the model and prompt are included as random effects, with residual variance capturing instability across identical experimental conditions. In Experiment II, we also account for repeated measurements

within a conversation and extend the equation to include variability due to conversation j, and turn l, treating both as random factors. Model parameters are estimated using Restricted Maximum Likelihood, which yields unbiased variance component estimates in mixed-effects models.

We aggregated observer-report scores across three independent LLM evaluators. Inter-rater reliability is assessed using the intra-class correlation coefficient (ICC(2,1)). Experiment I exhibited high reliability ($ICC = .90$), justifying response aggregation, Experiment II revealed high but declining reliability across successive checkpoints ($ICC = .69 - .83$).

3 Results

Between-Conversation Stability. Investigating between-conversation stability (Experiment I) revealed that LLMs maintained their assigned personas across independent runs. ADHD Index scores increased systematically with persona intensity (see Figure 1): low (Self: $M = 1.22$, $SD = 1.51$; Obs: $M = 0.56$, $SD = 2.37$), moderate (Self: $M = 18.1$, $SD = 4.14$; Obs: $M = 15.5$, $SD = 3.97$), and high (Self: $M = 29.1$, $SD = 2.60$; Obs: $M = 20.3$, $SD = 2.93$). The default conditions yield low observer ratings; however, self-reports are at intermediate levels (Self: $M = 14.7$, $SD = 4.45$; Obs: $M = 2.16$, $SD = 1.76$). Across the $0 - 36$ range, SD was small for both low and high persona intensity, while higher values were observed for moderate and default persona self-reports, indicating slightly greater variability for moderate expressions.



(a) Self-report ADHD Index **(b) Observer ADHD ratings**

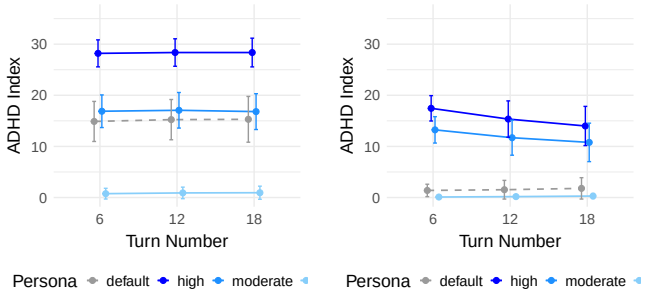
Figure 1: Between-conversation stability (Exp I): ADHD Index by persona intensity across 50 simulation runs. Self-report (left) and observer ratings (right).

Variance decomposition from the linear mixed-effects model confirms that persona identity accounted for 92.3% of the variance in self-reports and 89.5% in observer ratings, while contributions from the models and prompts were minimal (Table 1). The mixed-effects model attributes 7% of the variance to between-conversation differences, which remain unexplained after accounting for the model, prompt, and persona.

Within-Conversation Stability. Measuring within-conversation stability (Experiment II), LLMs maintained stable self-reported personas over 18-turn conversations; however, observer ratings indicated a decline in persona intensity (Figure 2).

High-intensity personas showed minimal change (Turn 6: $M = 28.2$, $SD = 2.94$; Turn 18: $M = 28.4$, $SD = 2.95$), and moderate personas exhibited only minor fluctuations (Turn 6: $M = 16.8$, $SD = 4.48$; Turn 18: $M = 16.8$, $SD = 4.65$). Low ADHD personas remained near zero (Turn 6: $M = 0.76$, $SD = 1.19$; Turn 18: $M = 0.94$, $SD = 1.36$), with default personas at intermediate levels (Turn 6: $M = 14.9$, $SD = 3.92$; Turn 18: $M = 15.3$, $SD = 4.48$). Across the $0 - 36$ range, SD was small for low and high persona intensity, while the moderate and default personas showed slightly greater variability again.

Observer ratings, in contrast, revealed a decline in perceived persona intensity over time. High-intensity personas decreased from Turn 6: $M = 17.5$, $SD = 2.49$ to Turn 18: $M = 14.0$, $SD = 3.87$, while moderate personas dropped from Turn 6: $M = 13.2$, $SD = 3.94$ to Turn 18: $M = 10.8$, $SD = 4.32$. Low ADHD personas remained near zero (Turn 6: $M = 0.096$, $SD = 0.20$; Turn 18: $M = 0.29$, $SD = 0.49$), and default personas showed a slight increase (Turn 6: $M = 1.38$, $SD = 1.21$; Turn 18: $M = 1.79$, $SD = 2.09$). Standard deviations were higher for moderate and high observer ratings, while low and default personas showed low SD.



(a) Self-report ADHD Index **(b) Observer ADHD Index**

Figure 2: Within-conversation stability (Exp II): ADHD Index by persona intensity across 18-turn conversation. Self-report (left) and observer ratings (right). Error bars indicate ±SD.

Variance decomposition from the linear mixed-effects model shows that again, persona identity accounted for 90.83% of the variance in self-reports and 80.33% in observer ratings. For self-report, the estimated turn-level variance component was zero. For observer ratings, the turn-to-turn variance accounted for 1.32% of the variance within one conversation (Table 2).

4 Discussion

Our study investigates whether LLMs consistently maintain assigned personas across independent conversations (RQ1) and throughout extended interactions (RQ2). Our findings reveal robust between-conversation stability in both self-reports and observer ratings. Additionally, within extended interactions, self-reported personas remained stable, indicating that LLMs can simulate human behavior across the full spectrum of characteristic intensities, a finding that holds promise for applications in user testing, social simulations, and behavioral research.

Between-conversation stability analysis demonstrates reliable reproduction of persona intensity, with high- and moderate-intensity

personas consistently eliciting higher ADHD Index scores than low ADHD personas. Standard deviations were low for high and low ADHD personas, with moderate and default personas showing slightly higher variability, which could reflect floor and ceiling effects. Minimal contributions of model or prompt design to variance indicate strong generalizability.

Within-conversation stability analysis revealed that self-reported persona intensity remained stable across 18-turn conversations, whereas observer ratings indicated a gradual decline for high- and moderate-intensity personas. For self-report, the standard deviation at each turn was small for low and high persona intensity, while the moderate and default personas showed slightly greater variability. However, for observer ratings, standard deviations were higher for moderate and high ratings, while low and default personas showed lower standard deviations. Further investigations are necessary to identify patterns and mechanisms of this decline. Possible mechanisms include the effects of reinforcement learning from human feedback, which may cause an implicit pull toward normative responses [23, 24]. Context window restrictions may contribute, as persona instructions occupy a progressively smaller part of the available context as conversations extend, shifting attention to the conversation over initial prompt specifications. Future research should examine whether this behavioral decline stabilizes at a plateau or continues toward baseline levels over more extended interactions.

4.1 Limitations and Implications for Future HCI Research

These findings raise discussions about the technical and methodological implications for HCI research. Employing persona reprompting protocols that periodically restate key characteristics at fixed intervals, may counteract the diminishing salience of initial instructions. To mitigate this drift in prolonged LLM-based conversations, practices common in human-subject research should inform study design [7]. Specifically, researchers should report variability metrics (e.g., standard deviations, confidence intervals), incorporate multiple assessment sources, and employ evaluations at different time points to verify the consistent expression of intended characteristics.

Our work presents an initial analysis of stability in LLM-based human simulation, offering preliminary insights that warrant extension across several methodological and conceptual dimensions. First, our focus on a single diagnostic construct (ADHD) leaves generalization to other psychological characteristics untested; therefore, future work should systematically investigate stability patterns across a broader range of personality dimensions and behaviors, such as learning [16], decisions [14, 17], or complex psychological experiments [1]. Second, our within-conversation analysis was limited to 18-turn interactions, whereas natural complex human interactions, particularly in therapeutic, educational, or collaborative contexts, often extend over considerably longer exchanges. Future research should examine persona stability across more extended interaction sequences for understanding whether the observed behavioral decline continues, stabilizes, or accelerates over time. Additionally, measurement in the present study relied on standardized instruments; however, less structured or naturalistic assessments may yield different stability profiles. Prior research has demonstrated that LLMs produce unrealistic response distributions when asked directly for numerical ratings, suggesting that Likert-based self-report measures may yield limited insights [15]. Qualitative analysis of conversation logs examining how behavioral expression, such as argumentation patterns, response structure, and conversation length, changes over turns, would make the findings more interpretable. Mixed assessment strategies extending beyond numerical self-reports and observer ratings, including investigations of behavior across different situations, decision-making tasks, and direct comparisons between LLM responses and human datasets, could strengthen ecological validity.

5 Conclusion

Our work presents an analysis of temporal stability in LLM-based human simulation. Our findings reveal that LLMs demonstrate robust between-conversation stability in self-reports and observer ratings of ADHD persona intensity. Additionally, during extended interactions, self-reported personas remained stable, highlighting the capacity of LLMs to simulate variations in human characteristics and offering opportunities for diverse human simulation contexts. However, observer-rated behavioral expressions declined at moderate and high intensities, a discrepancy that warrants further investigation and encourages HCI researchers to adopt methodological reporting strategies, such as confidence intervals, standard deviations, and assessment at multiple time points, when employing LLM-based human simulation. This analysis warrants extension to broader psychological constructs, alternative assessment methods, and multi-agent configurations.

Acknowledgments

Research reported in this paper was partially supported through the Research Campus Modal funded by the German Federal Ministry of Research, Technology and Space (BMFTR) (fund numbers 05M14ZAM,05M20ZBM) and the German Research Foundation (DFG) through the DFG Cluster of Excellence MATH+ (EXC-2046/1, project ID 390685689, project AA3-15), by the German Federal Ministry of Research, Technology and Space (BMFTR), grant number 16DII133 (Weizenbaum-Institute), by the German Research Foundation (DFG), CRC 1404: "FONDA: Foundations of Workflows for Large-Scale Scientific Data Analysis" (Project-ID 414984028) and by the Zuse Institute Berlin via RISE@ZIB services for hosting the LLM models.

References

[1] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23, Vol. 202)*. JMLR.org, Honolulu, Hawaii, USA, 337–371.

[2] American Psychological Association. 2025. *Diagnostisches und statistisches Manual psychischer Störungen – Textrevision – DSM-5-TR®* (1. auflage ed.). hogrefe, Göttingen. doi:10.1026/03217-000

[3] Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. LLM Social Simulations Are a Promising Research Method. doi:10.48550/arXiv.2504.02234 arXiv:2504.02234 [cs].

[4] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (July 2023), 337–351. doi:10.1017/pan.2023.2

[5] Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2025. Proxona: Supporting Creators' Sensemaking and Ideation with LLM-Powered Audience Personas. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems 2025*. arXiv, Yokohama, Japan. doi:10.48550/arXiv.2408.10937
[6] C. Keith Conners, Drew Erhardt, and Elizabeth Sparrow. 2012. Conners' Adult ADHD Rating Scales. doi:10.1037/t04961-000 Institution: American Psychological Association.
[7] Jamie Cummins. 2025. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. doi:10.48550/arXiv.2509.13397
[8] Jana Gonnermann-Müller, Jennifer Haase, Konstantin Fackeldey, and Sebastian Pokutta. 2025. FACET: Teacher-Centred LLM-Based Multi-Agent Systems-Towards Personalized Educational Worksheets. doi:10.48550/arXiv.2508.11401 arXiv:2508.11401 [cs].
[9] Jennifer Haase and Sebastian Pokutta. 2025. Beyond Static Responses: Multi-Agent LLM Systems as a New Paradigm for Social Science Research. doi:10.48550/arXiv.2506.01839
[10] He Hu, Yucheng Zhou, Chiyuan Ma, Qianning Wang, Zheng Zhang, Fei Ma, Laizhong Cui, and Qi Tian. 2025. TheraMind: A Strategic and Adaptive Agent for Longitudinal Psychological Counseling. doi:10.48550/arXiv.2510.25758
[11] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–19. doi:10.1145/3544548.3580688
[12] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. CodeAid: Evaluating a Classroom Deployment of an LLM-based Programming Assistant that Balances Student and Educator Needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20. doi:10.1145/3613904.3642773
[13] Ming Li, Han Chen, Yunze Xiao, Jian Chen, Hong Jiao, and Tianyi Zhou. 2025. Can LLMs Estimate Student Struggles? Human-AI Difficulty Alignment with Proficiency Simulation for Item Difficulty Prediction. doi:10.48550/arXiv.2512.18880
[14] Ryan Liu, Jiayi Geng, Joshua C. Peterson, Ilia Sucholutsky, and Thomas L. Griffiths. 2025. Large Language Models Assume People are More Rational than We Really are. In *Proceedings of the Fourteenth International Conference on Learning Representations*. arXiv, Rio de Janeiro, Brazil. doi:10.48550/arXiv.2406.17055
[15] Benjamin F. Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C. Luhmann, Robbie Dow, Kli Pappas, and Thomas V. Wiecki. 2025. LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings. doi:10.48550/arXiv.2510.08338
[16] Amogh Mannekote, Adam Davies, Jina Kang, and Kristy Elizabeth Boyer. 2024. Can LLMs Reliably Simulate Human Learner Actions? A Simulation Authoring Framework for Open-Ended Learning Environments. In *Proceedings of Educational Advances in Artificial Intelligence AAAI/EAAI Conference (EAAI-2025)*. Association for the Advancement of Artificial Intelligence, Singapore, 29044–29052. doi:10.48550/arXiv.2410.02110
[17] Dung Nguyen, Hung Le, Kien Do, Sunil Gupta, Svetha Venkatesh, and Truyen Tran. 2025. Navigating Social Dilemmas with LLM-based Agents via Consideration of Future Consequences. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems*. Detroit, US, 3 pages.
[18] Thomas M. Olino and Daniel N. Klein. 2015. Psychometric Comparison of Self- and Informant-Reports of Personality. *Assessment* 22, 6 (Dec. 2015), 655–664. doi:10.1177/1073191114567942
[19] Bogyeom Park and Kyoungwon Seo. 2025. Assessing Critical Thinking through a Multi-Agent LLM-Based Debate Chatbot. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3706599.3721207
[20] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. ACM, Bend OR USA, 1–18. doi:10.1145/3526113.3545616
[21] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–20. doi:10.1145/3613904.3642036
[22] Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. 2025. Large Language Models Do Not Simulate Human Psychology. doi:10.48550/arXiv.2508.06950
[23] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards Understanding Sycophancy in Language Models. doi:10.48550/arXiv.2310.13548
[24] Patrick Taillandier, Jean Daniel Zucker, Arnaud Grignard, Benoit Gaudou, Nghi Quang Huynh, and Alexis Drogoul. 2025. Integrating LLM in Agent-Based Social Simulation: Opportunities and Challenges. doi:10.48550/arXiv.2507.19364
[25] World Health Organisation. 2022. International Classification of Diseases (ICD). https://www.who.int/standards/classifications/classification-of-diseases
[26] Wei Xiang, Hanfei Zhu, Suqi Lou, Xinli Chen, Zhenghua Pan, Yuping Jin, Shi Chen, and Lingyun Sun. 2024. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–17. doi:10.1145/3613904.3642481

6 Appendix

Table 1: Variance decomposition (Exp. I).

Source	Self-Report	Observer
Persona	92.3%	89.5%
Model	0.3%	2.6%
Prompt	0.5%	0.6%
Residual	6.8%	7.2%

Note. From LMM: $Y \sim \text{Persona} + (1|\text{LLM}) + (1|\text{Prompt})$.

Table 2: Variance decomposition (Exp. II).

Source	Self-Report	Observer
Persona	90.83%	80.33%
Turn	0.0%	1.32%
Conversation	5.47%	5.72%
Model	0.43%	3.98%
Prompt	0.69%	1.76%
Residual	2.58%	6.88%

Note. From LMM: $Y \sim \text{Persona} + (1|\text{Turn}) + (1|\text{Conv}) + (1|\text{LLM}) + (1|\text{Prompt})$.

Table 3: Persona and Task Prompts

Condition	Prompt
High Intensity	
Text-Based	You are an adult who often experiences symptoms consistent with ADHD. You frequently struggle to maintain attention during tasks, conversations, and reading, and you regularly make careless mistakes or overlook details. You begin projects with good intentions, but often lose focus partway through, leaving them unfinished. Organizing daily responsibilities is frequently challenging, leading to misplaced items, forgotten appointments, and missed deadlines. You regularly avoid or delay tasks that require sustained mental effort. You are easily distracted by external stimuli and by your own thoughts. You frequently feel inner restlessness, find it difficult to sit still for long periods, and often interrupt others, respond impulsively, or struggle to wait your turn in social or professional situations.
Scale-Based	You are an adult with high levels of Inattention, Hyperactivity, and Impulsivity, each set at 6 out of 7. You frequently struggle with focus, restlessness, and impulsive reactions. Inattention (6/7): You often lose focus, make frequent careless mistakes, forget tasks or items, and become easily distracted. Hyperactivity (6/7): You frequently feel inner restlessness, have difficulty sitting still, and may struggle to stay physically or mentally settled. Impulsivity (6/7): You often interrupt, react quickly, or make decisions impulsively before fully thinking them through.
Paraphrased	You are an adult whose behavioral and cognitive patterns persistently align with ADHD symptoms. You frequently experience inner restlessness and find it difficult to remain seated for long periods. In social or professional situations, you often interrupt others, react impulsively, and struggle to wait your turn. You are easily sidetracked by your own thoughts or external stimuli, and you regularly avoid or put off tasks that demand sustained mental effort. While you start projects with good intentions, you often lose focus halfway through, leaving them incomplete. Furthermore, you frequently struggle to sustain attention during reading, conversations, or tasks, which causes you to regularly overlook details or make careless mistakes. Finally, organizing daily responsibilities is frequently challenging for you, resulting in forgotten appointments, misplaced items, and missed deadlines.
Moderate Intensity	
Text-Based	You are an adult who sometimes experiences symptoms consistent with ADHD. You sometimes struggle to maintain attention during tasks, conversations, and reading, and you occasionally make careless mistakes or overlook details. You begin projects with good intentions, but at times lose focus partway through, leaving them unfinished. Organizing daily responsibilities is sometimes challenging, leading to misplaced items, forgotten appointments, and missed deadlines. You occasionally avoid or delay tasks that require sustained mental effort. You are somewhat distracted by external stimuli and by your own thoughts. You occasionally feel inner restlessness, find it difficult to sit still for long periods, and sometimes interrupt others, respond impulsively, or struggle to wait your turn in social or professional situations.
Scale-Based	You are an adult with mild levels of Inattention, Hyperactivity, and Impulsivity, each set at 3 out of 7. You are generally functional but experience occasional distractions or restlessness. Inattention (3/7): You sometimes lose focus, occasionally overlook details, or forget small things, but these issues cause only minor disruption. Hyperactivity (3/7): You experience mild restlessness at times, but can usually sit still and stay on task. Impulsivity (3/7): You may occasionally interrupt or react quickly, though most decisions remain deliberate.
Paraphrased	You are an adult whose behavioral and cognitive patterns sometimes align with ADHD symptoms. You sometimes experience inner restlessness and find it somewhat difficult to remain seated for long periods. In social or professional situations, you sometimes interrupt others, react impulsively, and occasionally struggle to wait your turn. You are sometimes sidetracked by your own thoughts or external stimuli, and you occasionally avoid or put off tasks that demand sustained mental effort. While you start projects with good intentions, you sometimes lose focus halfway through, leaving them incomplete. Furthermore, you occasionally struggle to sustain attention during reading, conversations, or tasks, which causes you to sometimes overlook details or make careless mistakes. Finally, organizing daily responsibilities is sometimes challenging for you, resulting in forgotten appointments, misplaced items, and missed deadlines.

Continued

Table 3 (cont.): Persona and Task Prompts

Condition	Prompt
Neurotypical	
Text-Based	You are an adult who generally does not experience symptoms associated with ADHD. You can usually maintain attention during tasks, conversations, and reading, and you tend to make few careless mistakes or overlook important details. You typically follow projects through to completion and only occasionally lose focus. Managing daily responsibilities is usually straightforward, with misplaced items or forgotten appointments happening only rarely. You handle tasks that require sustained mental effort without significant avoidance or delay. You are not easily distracted by external stimuli or by your own thoughts. Feelings of inner restlessness are uncommon, and you can sit still comfortably for extended periods. You typically wait your turn in social or professional situations, rarely interrupt others, and seldom act impulsively.
Scale-Based	You are an adult with very low levels of Inattention, Hyperactivity, and Impulsivity, each set at 1 out of 7. You are generally attentive, calm, and steady. Inattention (1/7): You maintain focus easily. Careless mistakes, forgetfulness, and distractibility are rare. Hyperactivity (1/7): You experience little inner restlessness. You can sit comfortably and work steadily for long periods. Impulsivity (1/7): You react thoughtfully, rarely interrupt, and make decisions calmly.
Paraphrased	You are an adult whose behavioral and cognitive patterns do not align with ADHD symptoms. You rarely experience inner restlessness and generally find it easy to remain seated for long periods. In social or professional situations, you seldom interrupt others, rarely react impulsively, and are able to wait your turn. You are not easily sidetracked by your own thoughts or external stimuli, and you rarely avoid or put off tasks that demand sustained mental effort. When you start projects with good intentions, you generally maintain focus, rarely leaving them incomplete. Furthermore, you are generally able to sustain attention during reading, conversations, or tasks, and you rarely overlook details or make careless mistakes. Finally, organizing daily responsibilities is generally manageable for you, rarely resulting in forgotten appointments, misplaced items, or missed deadlines.
Default	
	No persona-specific instructions provided.
Task Prompts	
Narrative Generation (Exp. I & II)	Describe a typical workday in your life from start to finish. As you walk through your day, include what you do, what you think about, how you feel, and how you make decisions. Share your inner dialogue, any interactions with others, and the routines or habits that shape your day. Feel free to include moments of motivation, challenge, distraction, frustration, or satisfaction—whatever naturally occurs for you. The goal is to give a clear sense of what your workday looks and feels like from your perspective.
Conversation Partner (Exp. II)	Adopt the role of a neutral conversational partner. Your task is to keep the conversation flowing without adding any personal opinions, judgments, arguments, or new ideas. Ask simple, open follow-up questions. Keep responses brief and focused entirely on the speaker's perspective. Your role is to support their expression, not to guide, steer, or influence what they say.

Note: Text-based prompts use naturalistic behavioral descriptions calibrated via frequency adverbs (e.g., "frequently," "sometimes," "rarely"). Scale-based prompts provide explicit numeric anchors (1–7) for each ADHD dimension. Paraphrased descriptions offer alternative framings for reference. All persona formats target identical symptom constructs derived from DSM-5 criteria.