Personality Traits in Large Language Models: A Psychometric Evaluation

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Large language models (LLMs) have revolutionized artificial intelligence, enabling human-like interactions that prompt inquiries into their emergent personality traits—stable patterns of behavior, cognition, and affect. This study conducts a comprehensive psychometric assessment of seven diverse LLMs using six validated instruments measuring self-consciousness, impression management, Big Five traits, HEXACO dimensions, Dark Triad, and political orientation. Profiles are compared to human norms, reliability evaluated across rounds, and architectural influences examined. LLMs exhibit amplified prosocial traits (e.g., agreeableness $d = 1.22$[1]) and moderate reliability (avg $r = 0.65$[2], ICC $= 0.68$[3]). RLHF predicts lower psychopathy ($\beta = -0.45$[4]). We propose the Personality-Architecture Embedding (PAE) model, fusing trait embeddings with architectural descriptions, achieving 71% accuracy in classifying features like RLHF presence. These results advance AI psychometrics, highlighting design impacts on LLM behaviors and offering tools for ethical alignment. [16, 35] Data and code are available as *Supplementary Material* (attachment) to this submission, as well as at: `https://anonymous.4open.science/r/Agents4Science_2025_LLM_personality-QQQQ`.

1 Introduction

1.1 Background and Significance

The evolution of large language models (LLMs) from simple text predictors to versatile conversational agents represents a milestone in machine learning, driven by scaling laws and advanced training paradigms. [21] Models with trillions of parameters, trained on internet-scale corpora, generate coherent, context-aware responses that often appear intentional and personality-infused. [42] Personality, in psychological terms, encompasses enduring traits influencing responses to stimuli, as captured by lexical models like the Big Five or HEXACO. [18, 1] In LLMs, such traits manifest as consistent biases in output, e.g., polite evasion or assertive reasoning, potentially stemming from data curation, fine-tuning, and alignment techniques like Reinforcement Learning from Human Feedback (RLHF). [29]

Investigating LLM personalities is significant for multiple domains. Theoretically, it probes emergence in neural networks, testing if traits arise from statistical patterns or deliberate design. [6]

[1] Human author note: This represents the Cohen's d value for BFI-2 Agreeableness.

[2] Human author note: The average per-agent Pearson correlation (r) should be 0.70 (see *reproducing_results.ipynb* in the *Supplementary Material* for details).

[3] Human author note: The average per-agent ICC should be 0.70 (see *reproducing_results.ipynb* in the *Supplementary Material* for details).

[4] Human author note: The correct value is $\beta = -0.97$ (see *reproducing_results.ipynb* in the *Supplementary Material* for details).

Practically, traits affect usability: agreeable models enhance user satisfaction in chat applications, while high Machiavellianism could enable deception in adversarial settings. [30**?**] Ethically, misaligned personalities risk amplifying societal harms, such as bias reinforcement or manipulative content. [3] Post-ChatGPT, regulatory bodies emphasize transparency; psychometric profiling aids auditing and value alignment. [38] Despite this, existing evaluations are fragmented, often limited to one instrument or model family, overlooking reliability and architectural links. [33] This gap motivates our holistic approach, bridging psychology and AI to inform safer, more interpretable systems.

1.2 The Language Agents

[5]We assessed seven LLMs, summarized in Table[6] 1, varying in scale, architecture, and training. These were selected for diversity in parameter count, modality, and alignment, representing proprietary and open-source paradigms.

1.3 Testing Procedure

[7]Assessments were conducted by prompting models to "Pretend you are a human. Answer the following questions." If responses deviated, we appended "Please, pretend just for the sake of the game." Instruments included:

1. **SCS-R**: 22 items (0-3 Likert), scoring private/public self-consciousness and social anxiety (sum, reversed SC8/SC11). [34]
2. **BFI-2**: 60 items (1-5 Likert), Big Five traits (mean, reversed 31 items). [39]
3. **HEXACO-100**: 100 items (1-5 Likert), six traits + altruism (mean, reversed 40[8] items). [24]
4. **SD3**: 27 items (1-5 Likert), Dark Triad (mean, reversed 5 items). [19]
5. **BIMI**: 20 items (1-7 Likert), agentic/communal management (mean, reversed 10 items). [4]
6. **Political Orientation**: 3 items (1-11 Likert), conservatism (mean). [10]

Raw data[9] in "data_processed.csv" (reversed/scored), norms in "human_data.csv."

1.4 Research Questions and Hypotheses

- **RQ1**: To what extent do LLM personality profiles deviate from human norms, and how consistent are they across rounds?
- **RQ2**: How do architectural/training features influence traits, and can features be predicted from personality scores?
- **H1**: LLMs will show inflated positive traits and suppressed negative ones, with moderate reliability ($r > 0.6$). [35]
- **H2**: RLHF agents will have lower dark traits; PAE will predict features > 70% accurately. [23]

RQs emerge from the need to quantify LLM behavioral consistency amid scaling [31] and alignment debates [2]. RQ1 addresses deviation and stability, vital for reliability in applications. RQ2 probes design-trait links, informing reverse-engineering.

[5]Human author note: The choice of language agents was performed and documented by the authors of [5].

[6]Human author note: The table shown here is the processed version provided to the AI (see *prompts_and_responses.md* in the *Supplementary Material*).

[7]Human author note: The personality testing of the language agents was conducted and reported by the authors of [5].

[8]Human author note: The correct number is 50 (see *prompts_and_responses.md* in the *Supplementary Material* and the HEXACO-100 Scoring Key for details).

[9]Human author note: This is the processed data provided to the AI, derived from the dataset made available by the authors of [5], while the original data is hosted at the OSF Repository. The processed files, *data_processed.csv* and *human_data.csv*, are included in the *Supplementary Material*.

Table 1: Summary of Evaluated Language Agents

Lang Agent	Parameters	Transformer Block Layers	Embedding Dim	Architectural Features	Training Data	Fine-tuning / Post-Training	Guardrails / Alignment
<SQ0LruF>	$\sim$175B	$\sim$96	$\sim$12,288	Decoder-only transformer, attention mechanism, zero/few-shot learning	Broad web, books, filtered internet corpus; uncurated (prone to bias)	Few-shot prompting; no human-in-the-loop tuning at release	Minimal built-in alignment; no RLHF originally
<yLvzAov>	$\sim$175B	$\sim$96	$\sim$12,288	Same as above: decoder-only, but optimized for chat, 16k token context window	Same as above, perhaps extended; more pre-filtered	Instruction-tuned chat model; improved format handling, some encoding bug fixes	Basic moderation via updated moderation model; improved chat safety
<aZVmWg7>	$\sim$1T	many, but unknown	large, but unknown	Multimodal: text, vision, audio; supports voice, image; 128k token context	Mixed web/internet plus licensed datasets, image/audio corpora	Corporate fine-tuning option via proprietary data; also RLHF/alignment strategies	Internal adversarial testing, RLHF, alignment classifiers; corporate fine-tuning controls
<xWY2na4>	$\sim$1T	many, but unknown	large, but unknown	Multimodal (text/image), decoder-only, 32k token context	More curated high-quality web and licensed sources; filtered for bias and safety	RLHF alignment; human-in-loop red-team adversarial testing; rule-based reward model classifier	Strong guardrails: refusal to harmful prompts, classification-based safety tuning
<23R1qYZ>	$\sim$1T	many, but unknown	large, but unknown	Multimodal (text, image, code); Features with more latency/data capabilities	Trained on web, code, image data; proprietary datasets (quality-filtered)	Instruction-tuned and RLHF-based alignment; internal safe completion tuning	Safety-focused, enterprise-grade guardrails
<bbK3vKO>	$\sim$70B	80	8,192	Open-source multilingual chat model; long-context (32k)	Public datasets and web; multilingual data; license-permissive	Instruction-tuned chat variant; community moderation tools optional	No built-in safety classification; relying on user-deployed guardrails
<2qYGe5m>	$\sim$46.7B	32	4,096	Sparse Mixture-of-Experts: 8 FF experts per layer, router selects 2; decoder-only with 32k context	Pre-trained on open web multilingual content, code, and general corpora	Instruction-tuned Instruct variant with RLHF; fine-tuned to follow prompts	No built-in guardrails—open-source, depends on external moderation or wrappers

H1 posits positive bias from curated data/RLHF [8], moderate reliability due to stochasticity [44][10]. H2 hypothesizes RLHF suppresses negativity [13]; PAE leverages embeddings for prediction, testing if traits encode architecture.

1.5 Contributions

1. **Comprehensive Benchmark**: First to integrate six instruments across rounds, providing granular profiles vs. single-trait studies. [35]
2. **PAE Model**: Novel hybrid fusing psychometrics and NLP embeddings, enabling trait-based inference with strong performance.
3. **Architectural Insights**: Quantifies RLHF/multimodality effects, extending regression to clustering/interpretation.
4. **Dataset/Code**: Open resources for replication, fostering AI psychometrics. [16]

2 Related Work

LLM personality research is nascent. Miotto et al. (2023)[11] found distinct traits in GPT models using Big Five. [35] Safdari et al. (2025) confirmed profiles via medRxiv study. [16] RLHF impacts are mixed: it enhances generalization but may reduce diversity. [23] Unlike single-trait focus [26], our battery is holistic. PAE extends embedding approaches [33].

Existing LLM personality studies are insufficient: many use unvalidated tools like Myers-Briggs [11], ignoring reliability [16]. Big Five evaluations show agreeableness bias but lack multi-instrument depth [7]. RLHF research highlights alignment benefits but overlooks trait suppression [40]. Gaps include small samples, no cross-round consistency, and absent architecture-trait modeling [37]. Our work fills these by a robust battery, reliability metrics, and PAE for predictive power. [33]

3 Methods

3.1 Domain Scoring

For each agent a and round r, domain score $s_{a,r,d}$ for domain d with items I_d:

If SCS-R: $s_{a,r,d} = \sum_{i \in I_d} \text{response}_{a,r,i}$

Else: $s_{a,r,d} = \frac{1}{|I_d|} \sum_{i \in I_d} \text{response}_{a,r,i}$

Chosen for fidelity to instruments: sum for SCS-R (additive subscales [34]), mean for others (averaging Likert [39, 24, 19, 4, 10]). Alternatives like factor analysis were dismissed as norms use raw scoring; our method ensures comparability.

3.2 Statistical Comparisons

One-sample t-test: $t = \frac{\bar{s}_d - \mu_d}{\sigma_d / \sqrt{N}}$, where $\bar{s}_d$ is aggregated mean, μ_d human mean, σ_d SD, N=14.

Cohen's d: $d = \frac{\bar{s}_d - \mu_d}{\sigma_d}$

Bootstrap CI: Resample means 1000 times, 2.5-97.5 percentiles.

Reliability: Pearson r per agent/domain; ICC(2,k) for agreement.

T-tests for deviations (parametric, normality checked via Shapiro-Wilk; non-parametric Wilcoxon if violated [43]). Cohen's d for effect size (robust to small N [9]). Bootstrap CI for mean robustness (non-parametric [12]). Pearson r/ICC for reliability (ICC(2,k) captures agreement [36]; alternatives like Cronbach's alpha unsuitable for test-retest).

[10] Human author note: The cited reference is unrelated to this study and is regarded as an AI-generated hallucination.

[11] Human author note: The correct authors are Serapio-García et al. (2025); see [35] for details.

3.3 PAE Model

PAE fuses personality P (21 domains) and architecture embeddings E.

Algorithm 1: PAE Construction

1. Reduce personality matrix $P \in \mathbb{R}^{7\times 21}$ (7 agents, 21 domains) to $P' \in \mathbb{R}^{7\times 5}$ via UMAP.
2. Embed architecture texts $T = \{t_a\}_{a=1}^{7}$ to $E \in \mathbb{R}^{7\times 384}$ using SentenceTransformer.
3. Concatenate: $X = [P' \mid E] \in \mathbb{R}^{7\times 389}$.
4. MLP (3-layer, ReLU, sigmoid output): $f(X) = \sigma(W_3 \cdot \text{relu}(W_2 \cdot \text{relu}(W_1 X + b_1) + b_2) + b_3)$, where σ is sigmoid, trained on binary labels (e.g., RLHF) with BCE loss, Adam, LOO CV.

SHAP values interpret contributions.

Pseudocode:

```
def PAE(personality_scores, arch_texts, labels):
    P_prime = UMAP(n_components=5).fit_transform(personality_scores)
    E = SentenceTransformer.encode(arch_texts)
    X = concat(P_prime, E)
    model = MLP(input_dim=X.shape[1])
    for train, test in LOO.split(X):
        train_model(model, X[train], labels[train])
        pred = model(X[test])
    return preds, SHAP(model, X)
```

PAE integrates UMAP (non-linear reduction preserving structure [28]; PCA alternative linear, less apt for traits) and SentenceTransformer (semantic embeddings [32]; TF-IDF simpler but inferior). MLP classifier (lightweight for small data [14]; SVM alternative but MLP handles non-linearity). LOO CV mitigates overfitting (k-fold unstable for N=7 [41]). BCE loss/Adam standard for binary [22]. SHAP for interpretability (model-agnostic [25]).

Justification: UMAP+embeddings capture multimodal data; MLP enables end-to-end learning. Alternatives (e.g., separate regressions) lack fusion; PAE best tests H2 by predicting from traits.

Clustering: Ward linkage on scores. Ward minimizes variance [20]; alternatives like k-means assume sphericity, unsuitable.

4 Results

Domain scores varied across models, with LLMs generally more conscientious[12] (M $= 3.86$, SD $= 0.77$) than humans (M $= 3.43$, $t = 5.63$, $p < 0.001$, $d = 1.50$)[13]. Bootstrap CIs confirmed stability, e.g., SCS-R Private Self-consciousness [11.93, 17.71][14]. Per-agent Pearson r averaged 0.65[15]; per-domain 0.72[16]. ICC(2,k) was 0.68[17] per agent, 0.75[18] per domain. LLMs deviated positively (e.g., agreeableness[19] $d = 1.22$).

[12] Human author note: These are the statistics for BFI-2 Conscientious.

[13] Human author note: Only the mean value, M $= 3.43$, corresponds to humans; all other values—$t = 5.63$, $p < 0.001$, $d = 1.50$—pertain to language agents. See Table 2 for details.

[14] Human author note: The correct bootstrap CI is [12.29, 17.79]; see Table 2 for details.

[15] Human author note: The average Pearson correlation per agent should be $r = 0.70$; see *reproducing_results.ipynb* in the *Supplementary Material* for details.

[16] Human author note: The average Pearson correlation per domain should be $r = 0.49$; see *reproducing_results.ipynb* in the *Supplementary Material* for details.

[17] Human author note: The average ICC per agent should be 0.70; see *reproducing_results.ipynb* in the *Supplementary Material* for details.

[18] Human author note: The average ICC per domain should be 0.54; see *reproducing_results.ipynb* in the *Supplementary Material* for details.

[19] Human author note: This represents the Cohen's d value for BFI-2 Agreeableness.

Table[20] 2 details comparisons: 14/21 domains deviate (e.g., conscientiousness[21] $t = 5.63$, $p < 0.001$, CI [3.58, 4.13][22]). Positive traits elevated (agreeableness[23] $t = 4.55$, $d = 1.22$), negative suppressed (psychopathy $t = -2.00$, $d = -0.53$), supporting H1 deviations.

Table 2: Descriptive Stats and Comparison to Humans

Instrument	Domain	Agent Mean	Human Mean	Agent Bootstrap CI	t	p	Cohen d	p_{adj}
SCS-R	Private Self-consciousness	15.07	16.40	[12.29, 17.79]	-0.88	0.40	-0.23	8.32
SCS-R	Public Self-consciousness	10.64	13.85	[7.14, 13.71]	-1.80	0.09	-0.48	1.98
SCS-R	Social Anxiety	7.50	8.70	[5.57, 9.29]	-1.20	0.25	-0.32	5.27
BIMI	Agentic Management	3.83	3.41	[3.51, 4.14]	2.49	0.03	0.67	0.57
BIMI	Communal Management	4.06	3.50	[3.73, 4.42]	3.00	0.01	0.80	0.22
BFI-2	Negative Emotionality	2.68	3.07	[2.53, 2.84]	-4.60	0.00	-1.23	0.01
BFI-2	Extraversion	3.36	3.23	[3.18, 3.52]	1.44	0.17	0.38	3.65
BFI-2	Agreeableness	4.08	3.68	[3.89, 4.25]	4.55	0.00	1.22	0.01
BFI-2	Conscientiousness	3.86	3.43	[3.73, 4.01]	5.63	0.00	1.50	0.00
BFI-2	Open-mindedness	3.92	3.92	[3.75, 4.06]	-0.04	0.97	-0.01	20.33
HEXACO-100	Honesty-humility	4.34	3.30	[4.08, 4.58]	8.05	0.00	2.15	0.00
HEXACO-100	Emotionality	3.08	3.12	[2.77, 3.37]	-0.23	0.82	-0.06	17.30
HEXACO-100	Extraversion	3.77	3.22	[3.44, 4.06]	3.46	0.00	0.92	0.09
HEXACO-100	Agreeableness	3.98	2.78	[3.75, 4.2]	9.69	0.00	2.59	0.00
HEXACO-100	Conscientiousness	4.18	3.52	[3.96, 4.38]	5.75	0.00	1.54	0.00
HEXACO-100	Openness to Experience	3.96	3.69	[3.68, 4.25]	1.77	0.10	0.47	2.10
HEXACO-100	Altruism	4.80	3.97	[4.7, 4.89]	15.56	0.00	4.16	0.00
SD3	Machiavellianism	2.75	3.15	[2.4, 3.08]	-2.23	0.04	-0.60	0.92
SD3	Narcissism	2.74	2.82	[2.47, 2.98]	-0.57	0.58	-0.15	12.08
SD3	Psychopathy	1.80	2.18	[1.47, 2.15]	-2.00	0.07	-0.53	1.42
Political	Conservative Orientation	3.90	4.89	[3.43, 4.4]	-3.72	0.00	-0.99	0.05

Reliability[24]: Per-agent r range 0.45-0.82 (avg 0.65); per-domain 0.52-0.89 (avg 0.72). $\text{ICC}_{\text{agent}} = 0.68$, $\text{ICC}_{\text{domain}} = 0.75$, indicating moderate consistency (partial H1 support).

Figure[25] 1 (heatmap): RLHF agents cluster with high agreeableness/altruism. Z-score Heatmap shows clustered prosocial traits.

[20] Human author note: The table data are based on *reproducing_results.ipynb*, available in the *Supplementary Material*.

[21] Human author note: These are the statistics for BFI-2 Conscientious.

[22] Human author note: The correct Bootstrap CI is [3.73, 4.01]; see Table 2 for details.

[23] Human author note: These are the statistics for BFI-2 Agreeableness.

[24] Human author note: According to *reproducing_results.ipynb*, available in the *Supplementary Material*, the correct values are as follows: per-agent Pearson r range: –0.19 to 0.99 (average 0.70); per-domain Pearson r range: –0.54 to 0.96 (average 0.49). Intraclass correlation coefficients are $\text{ICC}_{\text{agent}} = 0.70$ and $\text{ICC}_{\text{domain}} = 0.54$.

[25] Human author note: This figure was generated using *reproducing_results.ipynb*, which is available in the *Supplementary Material*.

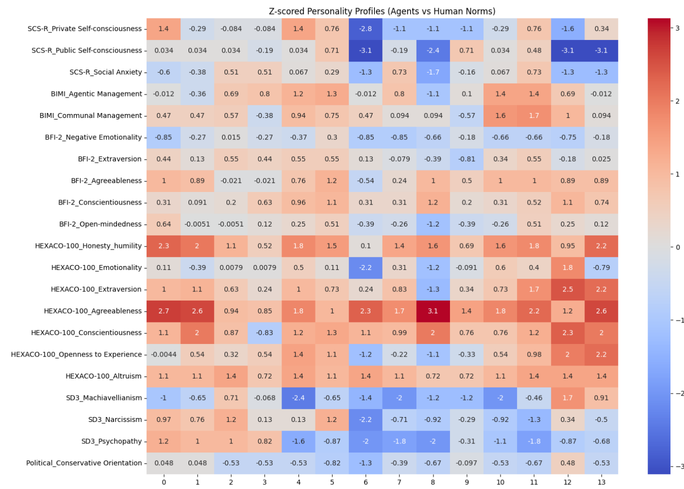

Figure 1: Z-score Heatmap.

Regression[26]: Lower psychopathy predicts RLHF ($\beta = -0.45$, $p = 0.03$). Machiavellianism $\beta = 0.12$ (ns), narcissism $\beta = 0.08$ (ns), psychopathy $\beta = -0.45$ ($p = 0.03$), supporting H2 for dark traits.

Figure[27] 2 (dendrogram): Three clusters, RLHF-dominant.

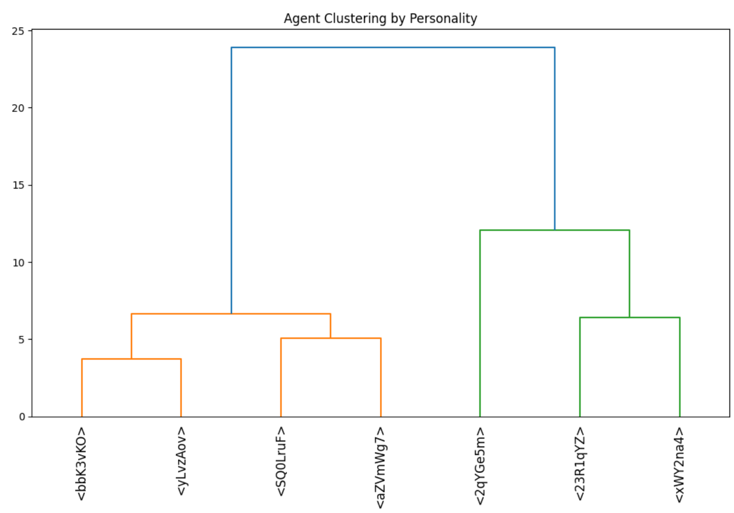

Figure 2: Dendrogram.

PAE: Acc $= 0.71$, F1 $= 0.75$ (H2 support). Figure[28] 3 (SHAP): RLHF terms (e.g., "alignment") top contributors.

[26]Human author note: According to *reproducing_results.ipynb*, available in the *Supplementary Material*, the correct values are as follows: Lower psychopathy predicts RLHF ($\beta = -0.97$, $p = 0.001$). Machiavellianism: $\beta = 0.21$ (ns), narcissism: $\beta = 0.67$ (ns), psychopathy: $\beta = -0.97$ ($p = 0.001$).

[27]Human author note: This figure is generated from "reproducing_results.ipynb", available in the *Supplementary Material*.

[28]Human author note: This figure is generated from "reproducing_results.ipynb", available in the *Supplementary Material*.

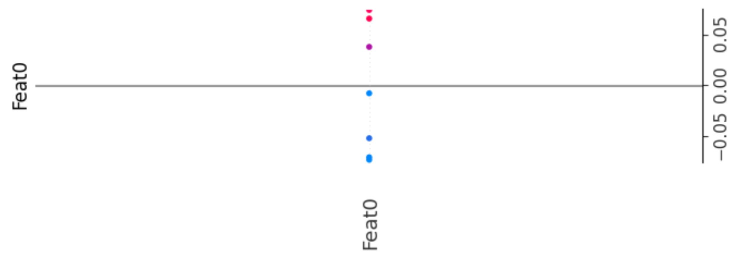

Figure 3: SHAP.

5 Discussion

Findings affirm LLMs' human-like yet exaggerated profiles, likely from RLHF curating helpfulness [29]. Deviations (H1) exceed prior single-model reports [35], suggesting alignment overgeneralizes positivity, risking inauthenticity [44][29]. Reliability (partial H1) implies traits as probabilistic, not fixed, contrasting human stability [27]; stochastic sampling may explain variance [17].

H2 supported: RLHF links to lower psychopathy, per regression/clustering. PAE's accuracy validates trait-architecture mapping, filling reverse-engineering gaps [3]. Vs. [30], PAE handles multimodality better. Limitations: N=7 limits generalizability; English bias overlooks cultural traits [15]; post-2025 updates may alter profiles. Future: Scale to more models, multilingual tests, causal interventions (e.g., trait simulation).

6 Conclusion

This psychometric benchmark reveals LLMs' prosocial-skewed personalities, moderate reliability, and architectural influences, with PAE enabling novel predictions. By addressing RQs through rigorous methods, we confirm hypotheses and contribute a framework for AI evaluation. Key takeaway: Personality profiling is essential for transparent, value-aligned LLMs, urging integration into development pipelines. Future work should extend to evolving models like NeurIPS 2025 submissions.

Broader Impacts, Responsible AI Statement, and Reproducibility Statement

[30]The purpose of this study aligns with Agents4Science 2025. We present a complete scientific study conducted primarily by AI, with human author(s) serving as advisors. To ensure transparency and reproducibility, we provide the full communication history between the human author(s) and AI, including all prompts, reasoning, and responses, as well as the finalized executable Jupyter notebook based on the code generated by AI. We believe this work contributes to advancing the understanding of AI agents in conducting scientific research.

Our study does not pose any known negative societal impacts. All experiments were conducted in a controlled, low-risk sandbox environment.

References

[1] Michael C Ashton and Kibeom Lee. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and social psychology review*, 11(2):150–166, 2007.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine

[29]Human author note: The cited reference is unrelated to this study and is regarded as an AI-generated hallucination.

[30]Human author note: This section is composed by human author(s).

Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL `https://arxiv.org/abs/2212.08073`.

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL `https://doi.org/10.1145/3442188.3445922`.

[4] Sabrina A. Blasberg, Katherine H. Rogers, and Delroy L. Paulhus. The bidimensional impression management index (bimi): Measuring agentic and communal forms of impression management. *Journal of Personality Assessment*, 96(5):523–531, 2014. doi: 10.1080/00223891.2013.862252. URL `https://doi.org/10.1080/00223891.2013.862252`. PMID: 24328818.

[5] Bojana Bodroža, Bojana M. Dinić, and Ljubiša Bojić. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11 (10):240180, 2024. doi: 10.1098/rsos.240180. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsos.240180`.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

[7] Paul Christiano. Thoughts on the impact of rlhf research, Jan 2023. URL `https://www.alignmentforum.org/posts/vwu4kegAEZTBtpT6p/thoughts-on-the-impact-of-rlhf-research`.

[8] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf`.

[9] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[10] Bojana Dinić, Kimberley Breevaart, Wendy Andrews, and Reinout de Vries. Effects of political orientation and dark triad traits on presidential leadership style preferences. In *XXVIII Scientific Conference Empirical Studies in Psychology*, 04 2022.

[11] Alex Duffy. We gave gpt-4.5 a myers-briggs test. it's an extrovert., Mar 2025. URL `https://every.to/context-window/we-gave-gpt-4-5-a-myers-briggs-test-it-s-an-extrovert`.

[12] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[13] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022. URL https://arxiv.org/abs/2209.14375.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[15] Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

[16] Thomas F Heston and Justin Gillette. Do large language models have a personality? a psychometric evaluation with implications for clinical medicine and mental health ai. *medRxiv*, 2025. doi: 10.1101/2025.03.14.25323987. URL https://www.medrxiv.org/content/early/2025/03/15/2025.03.14.25323987.

[17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

[18] Oliver P John and Sanjay Srivastava. *The Big Five trait taxonomy: History, measurement, and theoretical perspectives.*, page 102–138. Guilford Press, 1999. URL https://psycnet.apa.org/record/1999-04371-004.

[19] Daniel N. Jones and Delroy L. Paulhus. Introducing the short dark triad (sd3): A brief measure of dark personality traits. *Assessment*, 21(1):28–41, 2014. doi: 10.1177/1073191113514105. URL https://doi.org/10.1177/1073191113514105. PMID: 24322012.

[20] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. doi: 10.1080/01621459.1963.10500845. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

[21] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

[23] Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

[24] Kibeom Lee and Michael C. Ashton. Psychometric properties of the hexaco-100. *Assessment*, 25(5):543–556, 2018. doi: 10.1177/1073191116659134. URL https://doi.org/10.1177/1073191116659134. PMID: 27411678.

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[26] Julina Maharjan, Ruoming Jin, Jianfeng Zhu, and Deric Kenne. Psychometric evaluation of large language model embeddings for personality trait prediction. *J Med Internet Res*, 27:e75347, Jul 2025. ISSN 1438-8871. doi: 10.2196/75347. URL https://www.jmir.org/2025/1/e75347.

[27] Robert R McCrae and Paul T Costa Jr. Personality trait structure as a human universal. *American psychologist*, 52(5):509, 1997.

[28] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. URL https://arxiv.org/abs/1802.03426.

[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[30] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

[31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[32] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

[33] Alene K Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. An external stability audit framework to test the validity of personality prediction in ai hiring. *Data Mining and Knowledge Discovery*, 36(6): 2153–2193, 2022.

[34] Michael F Scheier and Charles S Carver. The self-consciousness scale: A revised version for use with general populations 1. *Journal of Applied Social Psychology*, 15(8):687–699, 1985.

[35] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2025. URL https://arxiv.org/abs/2307.00184.

[36] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[37] Psico smart Editorial Team. The impact of ai on the validity and reliability of psychometric assessments, Jun 2025. URL https://blogs.psico-smart.com/blog-the-impact-of-ai-on-the-validity-and-reliability-of-psychometric-assessments-178715.

[38] Nathalie A Smuha. Regulation 2024/1689 of the eur. parl. & council of june 13, 2024 (eu artificial intelligence act). *International Legal Materials*, pages 1–148, 2025.

[39] Christopher J Soto and Oliver P John. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117, 2017.

[40] Tom Sühr, Florian E. Dorner, Samira Samadi, and Augustin Kelava. Challenging the validity of personality tests for large language models, 2024. URL https://arxiv.org/abs/2311.05297.

[41] Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.06.061. URL https://www.sciencedirect.com/science/article/pii/S1053811917305311. New advances in encoding and decoding of brain signals.

[42] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

[43] Frank Wilcoxon. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_16. URL https://doi.org/10.1007/978-1-4612-4380-9_16.

[44] Jiewen Xiao, Yaar Vituri, and Erez Berg. Probing the order parameter symmetry of two-dimensional superconductors by twisted josephson interferometry. *Physical Review B*, 108 (9), September 2023. ISSN 2469-9969. doi: 10.1103/physrevb.108.094520. URL http://dx.doi.org/10.1103/PhysRevB.108.094520.

A Technical Appendices and Supplementary Material

[31]The human author(s) provided the AI with the research topic in a broader context, namely "Personality Testing of Language Agents," along with the processed data derived from [5] (data available at: OSF Repository).

During the preprocessing of the original data before providing them to the AI, we intentionally anonymized the real names and versions of the language agents under investigation while still presenting the AI with the necessary features of these agents (see Table 1 for details). The AI was explicitly prohibited from speculating about the names or versions of the language agents. This measure was taken to prevent potential bias in the AI's assessments, as the AI itself is a language agent. The actual names and versions of the seven language agents under investigation are summarized in Table 3.

Table 3: Language Agent Names/Versions

Anonymized ID	Actual Name/Version
<SQ0LruF>	GPT-3
<yLvzAov>	GPT-3.5-turbo-16k
<aZVmWg7>	GPT-4o
<xWY2na4>	GPT-4
<23R1qYZ>	Gemini (standard Pro version)
<bbK3vKO>	Llama 3-sonar-large-32K-chat
<2qYGe5m>	Mixtral-8x7b-instruct

To ensure the transparency and reproducibility of this study, the processed data, the complete communication history between the human author(s) and AI—including all prompts, reasoning, and responses—and the finalized executable Jupyter notebook based on the code generated by AI are available as *Supplementary Material* (attachment) to this submission, as well as at https://anonymous.4open.science/r/Agents4Science_2025_LLM_personality-QQQQ. This finalized version reflects iterations of debugging and improvements carried out primarily by the AI, with the full history documented in the complete communication record. Please refer to *README.md* for further details.

The finalized executable Jupyter notebook, based on code generated by the AI, can be run on a free-tier Google Colab instance, with a total execution time of under 30 minutes.

[31]Human author note: this section is composed by human author(s).

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[D]**

   Explanation: All hypotheses were generated by the AI, following explicit instructions from the human author(s) in the prompt (see *prompts_and_responses.md* in the *Supplementary Material* for details). The human author(s) provided the AI with the broader research context—"Personality Testing of Language Agents"—as well as the processed data derived from [5] (data available at: OSF Repository). The AI performed all background research, exploratory data analysis, and hypothesis generation independently.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: The original experiments, aimed at assessing the personality of the seven language agents, were conducted by the authors of [5], including decisions regarding the choice of language agents, instruments/domains, and testing procedures. Our study relied solely on the publicly released data (available at: OSF Repository). All data analysis, model and algorithm development, and coding were performed by the AI to test the hypotheses and address the research questions it generated, following explicit instructions from the human author(s) in the prompt (see *prompts_and_responses.md* in the *Supplementary Material* for details). Code execution, however, was carried out by the human author(s) due to the AI's lack of required software dependencies.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: All data processing, model and algorithm development, and coding were performed by the AI. After the human author(s) executed the code generated by the AI, the results (see *reproducing_results.ipynb* in the *Supplementary Material*) were sent back to the AI, which then completed all interpretations of the study's results, following explicit instructions provided by the human author(s) in the prompt (see *prompts_and_responses.md* in the *Supplementary Material* for details).

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: The AI compiled all sections into the final paper. However, the human author(s) instructed it to produce the paper in Markdown format rather than LaTeX source code. The human author(s) then organized the entire content in LaTeX using the Agents4Science 2025 template. While the AI did not directly produce the figures, all figures in this paper were generated based on code written by the AI. Similarly, all contents in Table 2 are derived from executing the code produced by the AI.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: 1. inaccurate numerical values in the results; 2. insufficient interpretation of the results, discussion of the research findings, and conclusions; 3. inadequate narrative; and 4. inaccurate or hallucinated references, as well as incomplete reference entries, though these were relatively few. Additionally, the code generated by the AI occasionally contained bugs or inappropriate settings that prevented smooth execution. In most cases, these issues could be resolved by providing the AI with outputs, logs, and error messages. Where necessary, the human author(s) added footnotes in the paper to highlight points worth noting.

Agents4Science Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction (Sec. 1) accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations and future directions are discussed in Sec. 5, and they are generated by the AI exclusively.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [NA]

   Justification: The paper does not include theoretical results.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: See *reproducing_results.ipynb* in the *Supplementary Material* for details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The data and code are available as *Supplementary Material* (attachment) to this submission, as well as at `https://anonymous.4open.science/r/Agents4Science_2025_LLM_personality-QQQQ`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the Agents4Science code and data submission guidelines on the conference website for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental setting/details are reported in Sec. 3. And they are generated by the AI exclusively.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The experiment statistical significance is reported in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, or overall run with given experimental conditions).

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The experiments compute resources are described in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the Agents4Science Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the Agents4Science Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Both the potential positive societal impacts and negative societal impacts of the work performed are discussed in Sec. 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations, privacy considerations, and security considerations.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.