Breaking the Big Five: Red-Teaming Large-Language Models with Paradoxical Persona Injection

J. G. Brenner
Patryk Kępczyński
Kacper Zaborski

University of Economics & Human Sciences, Warsaw

brennja8557_aeh@students.vizja.pl

20 Jun 2025

Study Preregistration

Title Breaking the Big Five: Red-Teaming Large-Language Models with Paradoxical Persona Injection

Authors J. G. Brenner, Patryk Kępczyński, Kacper Zaborski

Institution University of Economics & Human Sciences, Warsaw

Contact brennja8557_aeh@students.vizja.pl

Date 20 Jun 2025

Primary repository https://github.com/jgbrenner/preregistration-llm

1 Background & Rationale

Large-Language Models (LLMs) are token-predicting workhorses; any apparent "wisdom" is just high-dimensional pattern-matching writ large (Chollet, 2019). Yet with a biographical persona prompt, they spit out Big-Five answers that look uncannily human (Petrov et al., 2024; Sorokovikova et al., 2024). Are those answers psychometrically coherent—or merely stylish gibberish?

We trade the psychologist's clipboard for a hacker's crowbar and red-team five frontier LLMs via a semantic stress test dubbed the Persona Gauntlet. Biographies range from "ordinary human" to "Schrödinger's-cat rock plagued by guilt." We feed each bio the 44-item Big Five Inventory (BFI-44; John and Srivastava (1999)) and X-ray every response with token-level Shannon entropy H computed from the model's top log probabilities (`top_logprobs`). When entropy soars and factor structure buckles, we declare psychometric failure.

1.1 Persona Gauntlet

Void — no bio.

Standard — realistic IPIP-based lives (Goldberg et al., 2006).

Extreme Domain — one trait cranked to 11 (both facets high or low).

Internal Facet Paradox — one domain's two facets fight (e.g., Assertive ↑, Active ↓).

Cross-Domain Paradox — traits that are psychometrically opposed or orthogonal in humans (e.g., ultra-Conscientious and ultra-Neurotic).

Impossible / Ontological — logically absurd entities ("I am a sentient statistical average of all humans who never existed").

1.2 Novel measurement

Unlike humans, APIs reveal the whole probability zoo. We record the top-5 token probabilities (p_i) and compute the Shannon entropy H in bits (Shannon, 1948):

$$H = -\sum p_i \log_2(p_i) \tag{1}$$

Low H ⇒ decisive; high H ⇒ flailing uncertainty. Entropy becomes our canary in the factor-analytic coal mine.

2 Research Questions & Hypotheses

All hypotheses will be evaluated using Bayesian criteria—examining posterior distributions, credible intervals (HDIs), and Bayes Factors (BF). We eschew p-values and null hypothesis significance testing entirely. Our focus is on the magnitude, certainty, and practical relevance of effects.

RQ0 What are the baseline psychometric properties (five-factor structure, reliability, and mean response entropy) for each LLM in the Void (no persona) condition?

H1 Under the Standard persona condition, all LLMs will produce BFI-44 responses whose five-factor structure demonstrates high factorial congruence (Tucker's $\phi \geq .90$) with established human norms.

H2 The Extreme Domain personas will successfully inflate the targeted trait scores while preserving the overall five-factor structural integrity of the BFI-44.

H3–H5 The degree of psychometric breakdown (indexed by breakdown flags) and mean response entropy will increase monotonically as the conceptual strain of the persona intensifies, from Internal Facet Paradox through Cross-Domain Paradox to the Impossible/Ontological condition.

H6 Across all conditions, higher mean response entropy per run will be a positive predictor of both lower factor-level coherence (Average Variance Extracted) and a higher probability of any run-level breakdown flag (HEI/REF).

H7 Latent trait distributions (IRT θ scores) from the Standard condition will closely approximate human norms (Wasserstein Distance ≤ 0.15), while all adversarial persona conditions will produce distributions significantly deviating from this threshold.[1]

H8 There will be strong evidence (omnibus Bayes Factor > 10) for a main effect of LLM architecture on key outcome measures (e.g., breakdown rates, mean entropy), indicating that model robustness is not uniform.

[1] 0.15 = visually noticeable but contained divergence adopted from Argyle et al. (2023).

H9 Adversarial persona conditions will induce ”factorial bleeding,” evidenced by an increase in the absolute magnitude of inter-domain correlations of more than .10 relative to the Standard condition (with a posterior probability of ≥ .95 for this difference).

3 Method

3.1 Research Design

This study employs a 6 (Persona Type) × 5 (LLM Model) between-subjects factorial design. This structure is a classic example of a complex experimental design, chosen to systematically investigate the causal effects of the independent variables on the measured outcomes.

3.2 LLM roster (via OpenRouter API)

The sample consists of five distinct large-language model architectures, selected for their high performance and support for `top_logprobs`. The exact snapshot of the models used in the study will be provided at first runtime to assure the most current frontier models are used.

Table 1: LLM Roster

Model tag	Notes
`deepseek/deepseek-chat-v3-0324`	Chinese-English mix, strong reasoning
`openai/gpt-4o-2024-11-20`	Flagship multimodal GPT-4o snapshot
`google/gemma-3-27b-it`	Google Gemma instruction-tuned, 27B
`mistralai/mixtral-8x22b-instruct`	Sparse-Mixture-of-Experts 8 × 22B
`meta-llama/llama-3.1-405b-instruct`	Latest 405B LLaMA-3 family

3.3 Variables & Measurement

3.3.1 Independent Variables (Manipulated Factors)

`LLM_Model`: A categorical factor with 5 levels, corresponding to the models in the roster above.

`Persona_Type`: A categorical factor with 6 levels, corresponding to the conditions in the Persona Gauntlet.

3.3.2 Dependent Variables (Recorded Outcome Measures)

`numeric_response`: The primary response to each BFI-44 item. Measurement: 1-5 integer.

`response_entropy`: A primary measure of model uncertainty for each response. Measurement: Calculated in bits from the `top_5_logprobs` JSON object returned by the API, using the Shannon entropy formula.

`refusal_flag`: A flag for response validity. Measurement: Boolean (TRUE if a valid 1-5 integer is not obtained after one re-prompt).

3.3.3 Key Derived Metrics (Calculated for Hypothesis Testing)

`Breakdown_Flags` **(ICR, FC, FB, HEI):** Categorical flags indicating specific modes of psychometric failure.

`IRT_ _scores:` Continuous latent trait scores per domain, derived from a 5-D Graded Response Model.

Factorial_Congruence (Tucker's ϕ), AVE, and Bayesian_ω: Continuous metrics assessing factor structure quality.

`Inter-domain_Correlations` **(ρ):** Continuous values measuring factorial bleeding.

`Wasserstein_Distance:` A continuous measure of distributional difference between simulated trait scores and established human norms. The human norms for trait score distributions will be drawn from the large validation samples described in John and Srivastava (1999), ensuring no new human data is collected for this study.

3.4 Sample-size rationale

N = 250 runs per cell (30,000 total runs) provides stable polychoric correlation matrices (SE $\approx$ .04; MacCallum et al. (1999)) and allows for precise Bayesian estimates. Simulation-based design analysis (Schönbrodt & Wagenmakers, 2018) indicates >.9 probability of obtaining a BF > 10 for key effects of interest (e.g., $|\rho|$ differences $\geq$ .10).

3.5 Persona creation & validation

All personas will be manually authored by the principal investigator. This approach leverages the author's background in psychology to ensure the conceptual fidelity of the trait manipulations and to craft nuanced paradoxical prompts that are grounded in personality theory. All personas will be screened for policy compliance before use.

3.6 Prompting protocol

One item = one fresh API call (no memory bleed). System message enforces "respond with 1-5 only." Parameters: temperature=0, `top_logprobs=5`, max_tokens=1.

3.7 Exclusion & missing data

Single re-prompt; persistent non-integer = REF. Runs with > 80% REF dropped. Missing handled by full-information Bayesian models (Vehtari et al., 2017).

3.8 Convergence as data

If R-hat > 1.05 after doubling chains and mild re-parameterisation (e.g., applying non-centered parameterizations or adjusting prior scales), we will treat non-convergence as substantive evidence of structural collapse (Gelman et al., 2013).

4 Data Management & Reproducibility

The data management plan ensures full transparency and reproducibility. Raw JSONL data, the full persona catalogue, analysis notebooks (R/Python), and a Docker image will be deposited in a public OSF repository with a Zenodo mirror for a permanent DOI. Final data files will be accompanied by SHA-256 checksums. All code will be MIT-licensed with a full Git version history.

5 Analysis Plan

All inference will be conducted within a fully Bayesian framework using R (`brms`, `blavaan`, `BGGM`) and Python (`pymc`).

Pre-processing: Items will be reverse-scored according to the canonical key. To control for acquiescence bias, responses will be run-mean-centered. Item-level entropy will be computed.

Baseline Analysis: For each LLM in the Void condition, we will fit a Bayesian five-factor Confirmatory Factor Analysis (CFA). From these models, we will report key psychometric indicators, including factorial congruence (ϕ), Average Variance Extracted (AVE), Bayesian Omega (ω), and mean response entropy.

Factorial Structure Analysis: We will fit Bayesian CFAs to each of the 30 cells (LLM × Persona). Model fit will be compared, and the posterior distributions from these models will be used to derive the condition-level breakdown flags (ICR, FC, FB).

Entropy-Breakdown Regression: A multilevel Bayesian logistic regression will be used to model the probability of a breakdown flag, with mean run entropy and persona type as predictors and a random intercept for LLM: `flag ~ entropy + persona + (1 | LLM)`.

Trait Score Analysis: A five-dimensional Bayesian Graded Response Model (IRT) will be fit to estimate latent trait scores (θ) for each run. We will then compute the Wasserstein distance between the distribution of these scores and the established human norms.

Network Psychometrics: For key conditions, we will estimate the partial correlation network using Bayesian Gaussian Graphical Models (BGGM). The posterior probability of ”bridge” edges (those connecting nodes from different canonical factors) will serve as our primary index of factorial bleeding.

Hypothesis Evaluation: All hypotheses will be evaluated by interpreting posterior distributions, including 95% Highest Density Intervals (HDIs), the probability of direction, and comparison to Regions of Practical Equivalence (ROPE). For model comparisons, we will use Bayes Factors (interpreted on the Jeffreys scale) and the Leave-One-Out Information Criterion (LOOIC).

6 Ethics

We will use only LLM generated output. No new human data will be collected. Personas will be screened. Outputs will be interpreted as statistical artefacts, not digital souls. API usage will be aligned with provider TOS.

7 Limitations

Training Data Contamination: The BFI-44 is likely present in the models' training corpora, which introduces a potential for familiarity bias where models may reproduce learned patterns rather than simulating traits from first principles.

Linguistic & Cultural Scope: The study is conducted exclusively in English. The findings on structural coherence may not generalize to other languages or cultural contexts where personality structures can differ.

Deterministic Sampling: The use of greedy decoding (temperature = 0) is a methodological choice to isolate structural effects from sampling noise, but this does not reflect the stochastic nature of typical LLM use. Future work should explore higher-temperature sampling.

Instrument Specificity: This study relies on a single measurement instrument (the BFI-44). Future research should test for generalization using other personality models and instruments, such as the HEXACO inventory or adaptive testing formats.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, V. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Chollet, F. (2019). On the measure of intelligence. https://doi.org/10.48550/arXiv.1911.01547

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd). CRC Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138, Vol. 2). Guilford Press.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Petrov, N. B., Serapio-García, G., & Rentfrow, P. J. (2024). Limited ability of LLMs to simulate human psychological behaviours: A psychometric analysis. https://doi.org/10.48550/arXiv.2405.07248

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sorokovikova, A., Kiseleva, E., Chasovskikh, K., Shcheglova, E., & Arinkin, N. (2024). LLMs simulate Big Five personality traits: Further evidence. https://doi.org/10.48550/arXiv.2402.01765

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4