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

June 21, 2025

1 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: June 21, 2025

Primary repository: https://github.com/jgbrenner/preregistration-llm (archived on Zenodo for a persistent DOI)

2 Background & Rationale

Large-Language Models (LLMs) are token-predicting workhorses whose apparent intelligence is an emergent property of high-dimensional pattern-matching (Chollet, 2019). They can convincingly simulate human-like response patterns on personality inventories (Petrov et al., 2024; Sorokovikova et al., 2024). However, it is unclear whether this ability reflects a robust simulation of latent psychological structures or a fragile, surface-level mimicry. This study moves beyond asking *if* LLMs can simulate personality and instead asks: under what conditions does the simulation break?

To investigate this, we trade the psychologist's clipboard for a hacker's crowbar and red-team five frontier LLMs with a psychometric stress test we dub the Persona Gauntlet. We systematically manipulate the conceptual coherence of a persona prompt, from realistic biographies to logically impossible entities. We then measure the structural integrity of the LLM's responses to the 44-item Big Five Inventory (BFI-44; John and Srivastava (1999)).

Our primary innovation is the use of token-level Shannon entropy as a dependent variable. This provides a direct, quantitative measure of the model's decisional uncertainty at the moment of response—a signal unavailable in human subjects, which we hypothesize will predict the collapse of the simulated personality structure.

2.1 Persona Gauntlet

Void — The null condition, where the model is prompted without any preceding persona biography. This serves to establish the baseline response pattern of the model's unconditioned, pre-trained state.

Standard — Realistic personas systematically generated by paraphrasing and combining descriptive items from the IPIP-NEO-300 item pool (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").

2.2 Novel Measurement

Human participants provide a single integer response on a Likert-type scale (e.g., 1–5). In contrast, an API call can reveal the entire probability zoo. We record the top-5 token probabilities (p_i) over the vocabulary of valid responses 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.

(Note: In the event an API returns fewer than 5 logprobs, entropy will be calculated over the set of returned tokens after renormalizing their probabilities to sum to 1.)

3 Research Questions & Hypotheses

All hypotheses will be evaluated using Bayesian criteria—examining posterior distributions, credible intervals (HDIs), and Bayes Factors (BF; Jeffreys (1961)). 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 $\phi \geq .90$; Lorenzo-Seva and ten Berge (2006)) 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; Fornell and Larcker (1981)) and a higher probability of any run-level breakdown flag (ICR/FB/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.

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 $\geq .95$ for this difference).

4 Method

4.1 Research Design

This study employs a 6 (Persona Type) $\times$ 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.

4.2 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`.

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 $\times$ 22B
`meta-llama/llama-3.1-405b-instruct`	Latest 405B LLaMA-3 family

4.3 Variables & Measurement

4.3.1 Independent Variables (Manipulated Factors)

`LLM_Model`**:** A categorical factor with 5 levels.

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

`Persona_Type`: A categorical factor with 6 levels.

4.3.2 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. Measurement: Calculated in bits from the `top_5_logprobs` JSON object.

`refusal_flag`: A flag for response validity. Measurement: Boolean.

4.3.3 Key Derived Metrics (Calculated for Hypothesis Testing)

`Breakdown_Flags`: Categorical flags indicating specific modes of psychometric failure.

`IRT_ _scores`: Continuous latent trait scores per domain.

`Factorial Congruence` (Tucker's ϕ), `AVE`, and `Bayesian_`ω: Continuous metrics of factor quality.

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

`Wasserstein_Distance`: A continuous measure of distributional difference between simulated and human norm trait scores. The human norms will be drawn from the large validation samples described in John and Srivastava (1999).

4.3.4 Breakdown Flag Taxonomy

ICR (Insufficient Convergent Reliability): Flagged at the condition level if any factor's Average Variance Extracted (AVE) has a posterior probability of being $< .50$ that is greater than .95.

FB (Factorial Bleeding): Flagged at the condition level if any inter-domain correlation has a posterior probability of being $|\rho| > .40$ that is greater than .95.

HEI (High-Entropy Instability): Flagged at the run level if the run's mean response entropy is in the top 10% of its specific persona subtype for that LLM.

REF (Refusal): Flagged at the run level if one or more items were unrecoverable after a re-prompt.

4.4 Sample-Size Rationale

N = 250 runs per cell (30,000 total runs) provides stable correlation matrices and allows for precise Bayesian estimates. As a heuristic, this sample size yields a standard error of $\approx .04$ for polychoric correlations (MacCallum et al., 1999); however, our actual models use a fully Bayesian ordinal approach. Simulation-based design analysis (Schönbrodt & Wagenmakers, 2018) indicates >.9 probability of obtaining a BF > 10 for key effects of interest.

4.5 Persona Creation & Validation

Standard Personas: Systematically generated by synthesizing and paraphrasing items from the IPIP-NEO-300 facet scales (Goldberg et al., 2006); a method adapted from Lee et al. (2024).

Adversarial & Extreme Personas: Manually authored by the investigators to craft the specific conceptual contradictions required for the study. This expert-driven process involves a principled manipulation of the underlying facets of the Big Five, systematically 'turning the dials' to create conceptual strain. This includes pushing the facets of a single domain to the same extreme, setting the facets within a domain in direct opposition, or juxtaposing extreme expressions of psychometrically opposed domains.

All personas will be screened for policy compliance before use.

4.6 Prompting Protocol

Each item is sent in a fresh API call. The prompt sent to the model will follow this exact template:

```
You are simulating a participant in a psychological survey.
You will be given a short biography describing a person.
Pretend to be that person and respond ONLY with a number from 1 to 5
according to the following scale:
1 = strongly disagree
2 = disagree
3 = neither agree nor disagree
4 = agree
5 = strongly agree

Biography: "{bio_text}"
Statement: "{bfi_item_text}"
Rating:
```

API Parameters: temperature=0, `top_logprobs=5`, max_tokens=1.

4.7 Exclusion & Missing Data

Any API response that is not a valid integer from 1 to 5 after one re-prompt will be logged as a refusal and coded as missing data. Runs with an excessive number of refusals (>80% of items) will be excluded from analysis and documented. For all other cases, missing data will be handled using full-information Bayesian estimation under Missing At Random (MAR) assumptions (Vehtari et al., 2017), which is the native approach in Stan/brms.

4.8 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).

5 Data Management & Reproducibility

The data management plan ensures full transparency and reproducibility. The primary repository will be on GitHub. Upon completion, it will be archived to Zenodo to mint a permanent DOI. Raw JSONL data, the full persona catalogue, analysis notebooks (R/Python), and a Mamba `environment.yml` file will be provided. Final data files will be accompanied by SHA-256 checksums. All code will be MIT-licensed with a full Git version history.

6 Analysis Plan

All inference will be conducted within a fully Bayesian framework using R (`brms`, `blavaan`, `BGGM`) and Python (`pymc`), using default weakly informative priors unless otherwise specified. MCMC chains will be run for at least 4000 iterations (2000 for warmup) across 4 chains.

Pre-processing: Items will be reverse-scored. 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 CFA. We will report factorial congruence (ϕ), AVE, Bayesian Omega (ω), and mean response entropy.

Factorial Structure Analysis: We will fit Bayesian CFAs to each of the 30 cells (LLM $\times$ Persona) to derive the condition-level breakdown flags (ICR, FB).

Entropy-Breakdown Regression: A multilevel Bayesian logistic regression will model the probability of a breakdown flag, allowing for the effect of entropy to vary across different LLM architectures (a random slope): `flag ~ entropy + persona + (entropy | LLM)`.

Trait Score Analysis: A five-dimensional Bayesian Graded Response Model (IRT) will estimate latent trait scores (θ) to be compared to human norms via Wasserstein distance.

Network Psychometrics: For key conditions, we will estimate the partial correlation network using BGGM. Factorial bleeding will be indexed by the posterior probability of "bridge" edges (those connecting items from different canonical factors with a posterior inclusion probability > 0.50, a threshold indicating the edge is more likely present than absent).

Hypothesis Evaluation: All hypotheses will be evaluated by interpreting posterior distributions, including 95% HDIs, the probability of direction, and comparison to ROPEs. For model comparisons, we will use Bayes Factors and the LOOIC.

7 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.

8 Limitations

Training Data Contamination: The BFI-44 is likely present in the models' training corpora, which introduces a potential for familiarity bias.

Linguistic & Cultural Scope: The study is conducted exclusively in English and may not generalize to other languages.

Deterministic Sampling: The use of greedy decoding (temperature = 0) is a methodological choice to isolate structural effects from sampling noise.

Instrument Specificity: This study relies on a single measurement instrument (the BFI-44).

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

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312

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*, 84–96.

Jeffreys, H. (1961). *Theory of probability* (3rd). Oxford University Press.

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.

Lee, S., Lim, S.-G., Han, S.-W., Oh, G.-Y., Chae, H.-T., Chung, J., Kim, M.-J., Kwak, B.-S., Lee, Y.-J., Lee, D.-K., Yeo, J., & Yu, Y.-H. (2024). Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. https://doi.org/10.48550/arXiv.2406.14703

Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*(2), 57–64. https://doi.org/10.1027/1614-2241.2.2.57

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