The Journal of Open Source Software

LLM Narrative Framework: A Tool for Reproducible Testing of Complex Narrative Systems

Peter J. Marko [1] and **Kenneth McRitchie** [1]

1 Independent Researcher

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 10 December 2025
Published: unpublished

Summary

Psychology has long struggled to empirically validate complex, holistic systems that produce narrative-based claims. To address this methodological gap, we developed the **LLM Narrative Framework**, an open-source, fully automated pipeline that uses Large Language Models (LLMs) as pattern-detection engines.

Our framework automates a rigorous "matching task" experimental design. It generates standardized narrative descriptions based on a system's rules, pairs them with ground-truth biographical data, and tasks blinded LLMs with identifying the correct matches against randomized controls. We designed the software to manage the entire research lifecycle: it handles data sourcing, generates factorial experimental designs, executes parallelized matching tasks via LLM APIs, and performs comprehensive statistical analysis. By treating the source system as an arbitrary algorithm, we provide a domain-agnostic tool for researchers to test the construct validity of any text-based framework—from personality typologies to sociological theories—at a scale that was previously impossible.

Statement of Need

In the wake of the replication crisis, social scientists face a difficult question: how can we apply quantitative rigor to qualitative or symbolic systems? Establishing construct validity in such frameworks has remained a stubborn challenge (Cronbach & Meehl, 1955). Traditional psychometrics require discrete, linear variables, while qualitative methods often lack scalability and statistical power.

The arrival of Large Language Models offers a solution. Recent research suggests LLMs can act as impartial "proxy raters" or pattern detectors (Argyle et al., 2023; Brown et al., 2020; Gilardi et al., 2023), leveraging their emergent reasoning capabilities (Kosinski, 2023; Wei et al., 2022). However, using them for rigorous scientific inquiry requires addressing the reproducibility crisis (Open Science Collaboration, 2015; The Turing Way Community, 2022). The **LLM Narrative Framework** addresses these needs by solving specific engineering challenges:

1. **Reproducibility:** LLMs are non-deterministic. Scientific inquiry requires strict versioning of prompts, parameters, and data.
2. **Scale:** Achieving statistical power requires thousands of high-context queries, which necessitates robust concurrency and error handling.
3. **Data Integrity:** Pipelines must ensure that the generation of stimuli (narratives) is rigorously blinded from the evaluation (matching).

We built the **LLM Narrative Framework** to solve these engineering challenges. It provides a standardized, "batteries-included" harness that allows researchers to define a source system (logic for generating profiles) and a target dataset (biographies), and then fully automates the testing process. While we demonstrate its utility using astrology as a high-noise "stress test"

(Carlson, 1985; Godbout, 2020), the framework is designed to be a general-purpose instrument for investigating weak signals in complex narrative data.

Architecture and Workflow

We organized the codebase (40,000+ lines of Python and PowerShell) into four primary architectural layers, designed to enforce separation of concerns and methodological transparency:

1. **Data Preparation Pipeline:** We implemented a deterministic ETL (Extract, Transform, Load) process to convert raw data into experimental stimuli. This layer includes:
   - **Automated Sourcing:** Scripts that fetch and structure raw biographical data.
   - **LLM-based Candidate Selection:** To ensure sample quality, we use LLMs to score subjects on metrics like historical eminence, applying a variance-based cutoff algorithm to optimize sample diversity.
   - **Text Neutralization:** A dedicated subsystem that automatically strips domain-specific jargon from descriptions, ensuring double-blind testing conditions.
2. **Experiment Orchestration:** The core engine manages the execution of complex factorial experiments.
   - **"Create - Check - Fix" Workflow:** We designed the system around a robust state-machine architecture. It creates experiments, audits them for completeness, and automatically repairs corrupted runs (handling API timeouts or parsing failures) without restarting from scratch.
   - **Configuration Archival:** To guarantee methodological reproducibility, every experiment automatically archives its exact configuration (`config.ini`) and manifest.
3. **LLM Integration:**
   - We abstracted API interactions (via OpenRouter) to support over 40 models (e.g., GPT-4, Claude, Llama, Gemini, DeepSeek).
   - We implemented resilient parsing logic to extract structured data ($k \times k$ matrices) from unstructured LLM narrative responses, allowing quantitative analysis of qualitative outputs.
4. **Analysis & Reporting:**
   - The framework automatically aggregates results into hierarchical CSVs (Replication $\rightarrow$ Experiment $\rightarrow$ Study).
   - It performs automated statistical testing (Three-Way Mixed ANOVA, Tukey HSD post-hoc, Benjamini-Hochberg FDR correction).
   - It generates publication-ready visualizations (boxplots, interaction plots) and calculates "lift" metrics to quantify performance relative to chance.

Validation

To ensure the framework serves as a sensitive and reliable instrument, we implemented a comprehensive test suite covering four pillars of validation:

1. **Unit Testing:** We use `pytest` to validate individual Python components.
2. **Integration Testing:** We verify end-to-end workflows in isolated sandboxes to ensure data integrity.
3. **Algorithm Validation:** We perform bit-for-bit verification of the personality assembly algorithms against a ground-truth expert system to ensure the stimuli are generated correctly.
4. **Statistical Validation:** We externally validated the analysis engine against **GraphPad Prism 10.6.1**. Our framework's output (p-values, F-statistics, effect sizes) matches the industry-standard software within a tolerance of ± 0.0001, ensuring it meets rigorous statistical standards for the behavioral sciences (Cohen, 1988; Dongen & Grootel, 2025; Jeffreys, 1961).

Availability

We are committed to open science principles. The full source code, documentation, and dataset are available on GitHub. The repository includes a comprehensive **Replication Guide** for reproducing our original study and a **Framework Manual** for researchers who wish to extend the tool to new domains.

Acknowledgements

We acknowledge the developers of the open-source libraries that made this work possible, particularly `pandas`, `scipy`, `statsmodels`, `pingouin`, and `seaborn`. We also thank the OpenRouter platform for facilitating access to a diverse range of LLM APIs.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Carlson, S. (1985). A double-blind test of astrology. *Nature*, *318*(6045), 419–425. https://doi.org/10.1038/318419a0

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281. https://doi.org/10.1037/h0040957

Dongen, N. van, & Grootel, L. van. (2025). Overview on the null hypothesis significance test: A systematic review on essay literature on its problems and solutions in present psychological science. *Meta-Psychology*, *9*, MP.2021.2927. https://doi.org/10.15626/MP.2021.2927

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Godbout, V. (2020). An automated matching test: Comparing astrological charts with biographies. *Correlation*, *32*(2), 13–41.

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

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *Proceedings of the National Academy of Sciences*, *120*(9), e2218926120. https://doi.org/10.1073/pnas.2218926120

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

The Turing Way Community. (2022). *The turing way: A handbook for reproducible, ethical and collaborative research*. https://doi.org/10.5281/zenodo.3233853

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Chowdhery, A., Narang, S., & Le, Q. V. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD