A Framework for Reproducible Testing of Complex Narrative Systems: A Case Study in Astrology

Peter J. Marko and Kenneth McRitchie

October 31, 2025

Abstract

Background: Psychology has struggled to empirically validate complex, holistic systems that produce narrative-based claims. This methodological gap highlights the need for new, more rigorous, and transparent research paradigms. **Objective:** This paper introduces and validates a novel, fully automated, and open-source framework for testing for weak signals in complex narratives. Using astrology as a challenging case study, we demonstrate a reproducible method for assessing the construct validity of a symbolic system against biographical data. **Methods:** We programmatically neutralized a library of astrological descriptions using a Large Language Model (LLM) to remove all esoteric terminology. We filtered a cohort of 10,707 individuals through a multi-stage, LLM-driven process to a final pool of 4,987 subjects, which we selected for eminence and psychological diversity. We then used independent LLMs as impartial arbiters to perform a series of matching tasks between biographies and personality descriptions. **Results:** Aggregate analysis revealed a statistically significant but practically negligible signal, creating a statistical tension that motivated a deeper, multi-level analysis. This decomposition uncovered two critical findings: 1) A "Goldilocks effect," where signal detection peaked at a medium level of task difficulty (k=10), and 2) Extreme heterogeneity across models, with signal detection capability varying by a factor of 575. This demonstrates that framework effectiveness is not universal but requires both a compatible model architecture and optimal task difficulty calibration. **Conclusion:** This study's primary contribution is a new, open-science paradigm for psychological research. By demonstrating its utility on a difficult and controversial topic, we provide a robust, methodologically reproducible, and scalable framework for future investigations into complex narrative systems.

Keywords: Psychology, astrology, large language models, computational social science, reproducibility, open science

Introduction

The replication crisis has spurred a fierce and ongoing debate within psychological science about methodological reform (van Dongen & van Grootel, 2025). A key challenge in this landscape is establishing the **construct validity** (i.e., whether a system measures what it claims to measure) of complex, holistic systems that generate narrative-based claims (Cronbach & Meehl, 1955). This paper introduces and validates the **LLM Narrative Framework**—an automated testing methodology that uses Large Language Models as pattern-detection engines to perform matching tasks, determining whether systematic signals in narrative descriptions can be detected at rates significantly greater than chance. Astrology serves as a prime example, where landmark empirical studies have faced significant methodological debate (Carlson, 1985; Eysenck & Nias, 1982; Ertel, 2009) and where comprehensive meta-analyses of quantitative research have consistently shown null results (Dean & Kelly, 2003). While modern "whole-chart" matching tests show promise (Currey, 2022; Godbout, 2020), even recent computational explorations have been limited by a reliance on opaque "black-box" tools and manual processes for

assessing semantic similarity (Marko, 2018). This history highlights the need for a fully automated, transparent, and scalable testing framework.

The advent of Large Language Models (LLMs) presents an opportunity to develop such a framework. Prior research on the construct validity of astrology has often employed matching tests, where judges attempt to pair biographical or psychological descriptions with their corresponding subjects (e.g., Carlson, 1985). LLMs, as powerful pattern-recognition engines (Google, 2024; Wei et al., 2022), are uniquely suited to automate this process. Unlike human judges, who are susceptible to cognitive biases, LLMs can be deployed as **agnostic arbiters**, executing a matching task at a massive scale. Recent research has shown that modern LLMs can meet or even exceed the reliability of human annotators for complex text-classification tasks (Gilardi et al., 2023) and can be used to simulate human samples for social science research (Argyle et al., 2023). This study introduces and validates such an LLM-based framework, using astrology as a challenging case study.

Our primary goal is to determine if a fully automated pipeline can serve as a sensitive instrument for detecting weak signals in complex, narrative-based claims. To this end, the study tests a single, core hypothesis: *that the LLM-based framework can distinguish between correctly mapped and randomly mapped personality descriptions at a rate significantly greater than chance*. While the successful detection of such a signal within the present case study of astrology has implications for that field, the broader contribution of this work is the validation of the methodology itself. The philosophical implications of using a non-conscious system to analyze subjective consciousness are taken up in a companion article (McRitchie & Marko, manuscript in preparation).

Methods

Tool Selection Across Pipeline Stages

The framework employs distinct LLMs for different stages of data preparation and evaluation, each selected to optimize performance, cost, and methodological independence (see Figures 1a and 1b at the end of this section for system architecture). For the LLM-based candidate selection stage, **LLM A (OpenAI's GPT-5)**[1] performed eminence scoring in batches of 100 subjects, leveraging its superior contextual understanding to score individuals against a fixed, absolute scale of historical impact. Subsequently, **LLM B (Anthropic's Claude 4.5 Sonnet)** generated OCEAN personality scores with a batch size of 50, chosen for its strong performance on nuanced psychological assessment tasks. For profile generation, **LLM C (Google's Gemini 2.5 Pro)** handled the neutralization of 149 astrological delineations, selected for its superior instruction-following capabilities and large context window. For the core matching task,

[1] The naming of the data generation models reflects the latest versions available at the time of the study. For provider details and release context, see Appendix C of the Supplementary Materials (Replication Guide).

seven independent evaluation models were deployed (see Table 2 in Experimental Design and Procedure for the complete experimental design). For the LLM-based candidate selection stage, **LLM A** used a calibrated eminence scoring prompt with fixed historical anchors (Jesus Christ = 100.0, Plato/Newton = 99.5, Einstein = 99.0) to establish an absolute scale, with explicit instructions to distinguish "lasting historical eminence" from "transient celebrity." **LLM B** used a structured prompt requesting Big Five personality traits rated on a 1.0-7.0 scale with JSON output format for automated parsing. Complete prompt texts for all LLM stages are available in the project repository.

To minimize potential data contamination, we selected evaluation models to be independent from data generation models where possible; we present a full discussion of this contamination risk in the Limitations section. We used a temperature of 0.0 for all evaluation models to maximize response consistency, a standard practice in LLM evaluation. While temperature=0.0 minimizes sampling variance, it does not eliminate it, as LLM APIs exhibit inherent non-determinism. Therefore, exact computational reproducibility is not achievable. Instead, our framework provides methodological reproducibility through transparent documentation and open-source code. We did not use randomization seeds in the original study, but future work could use them to enable exact replication of experimental stimuli. The consistency of our findings across 1,260 experiments demonstrates methodological stability despite this variance.

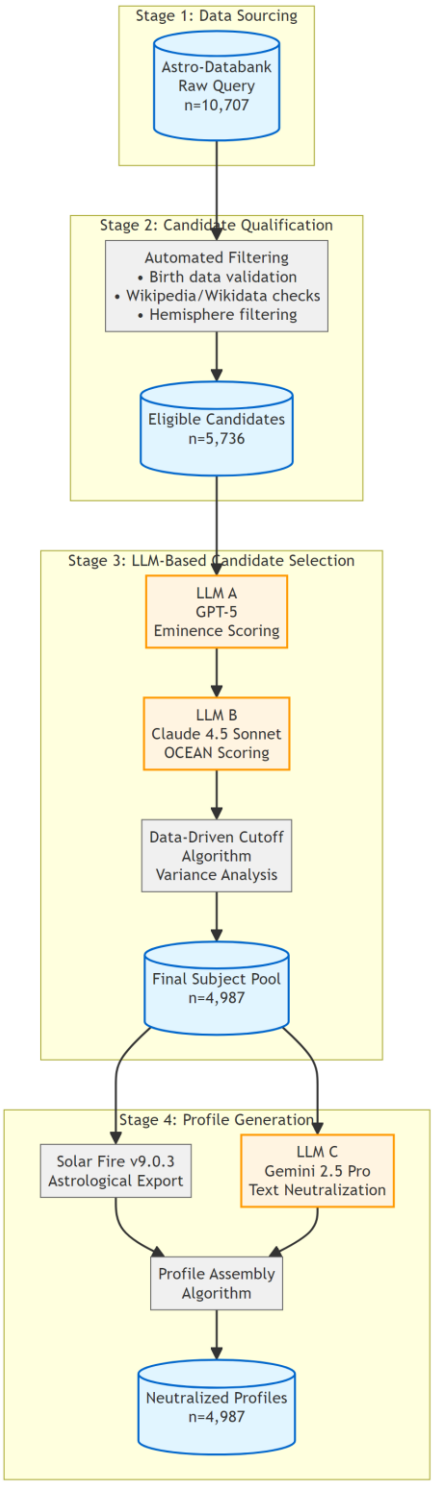

Figure 1a: The first four stages of the system architecture showing distinct LLM roles across the data preparation pipeline. LLM A (GPT-5) performs eminence scoring, LLM B (Claude 4.5 Sonnet) generates OCEAN scores, and LLM C (Gemini 2.5 Pro) neutralizes astrological text. This separation of roles is designed to minimize data contamination risk.

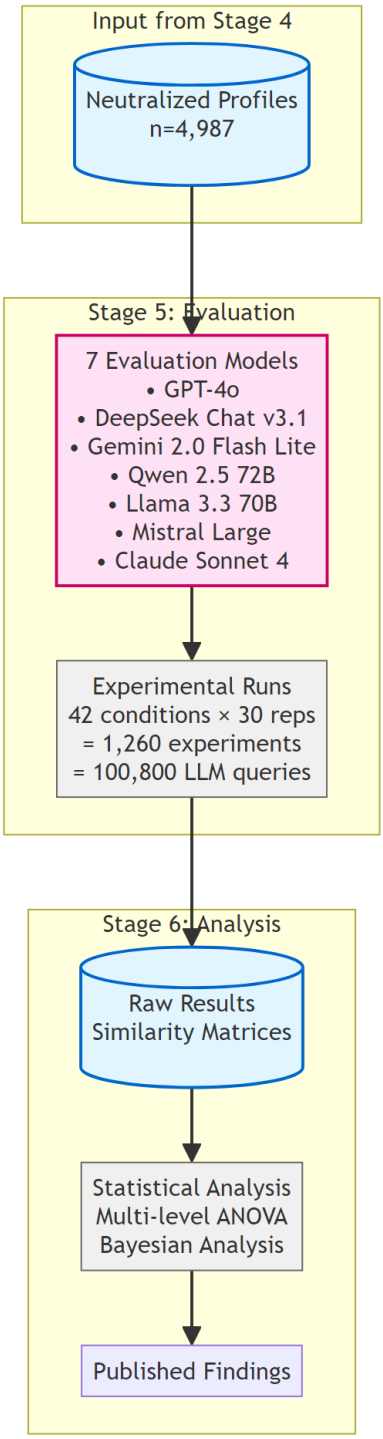

Figure 1b: The last two stages of the ystem architecture showing distinct LLM roles across the experiment & study workflow. Following the data the three preparation LLMs, seven independent evaluation models perform the matching task. This separation of roles is designed to minimize data contamination risk.

Sample Population

We designed the framework to support three distinct research paths. For **direct replication**, researchers can use the static data files included in the project's public repository. While the framework employs temperature=0.0 to minimize LLM response variance, exact computational reproducibility is not achievable due to inherent API non-determinism. Researchers should expect methodological reproducibility: statistically equivalent results using the same experimental design. For **methodological replication**, researchers can use the framework's automated tools to generate a fresh dataset from the live Astro-Databank (ADB) to test the robustness of the findings. Finally, for **conceptual replication**, researchers can modify the framework itself (e.g., by using a different LLM or analysis script) to extend the research.

We derived the final study sample from a multi-stage data preparation pipeline, as illustrated in Figure 2 below. This section provides a conceptual overview of the workflow; the **Supplementary Materials** (Replication Guide) contain a detailed, step-by-step guide for the entire pipeline (see Replication Guide in the online repository). In the first stage, **Data Sourcing**, we queried the Astro-Databank (ADB) for subjects based on three criteria: high-quality birth data (Rodden Rating 'A' or 'AA'), inclusion in the **Personal > Death** category to ensure the subject is deceased, and inclusion in the eminence category of **Notable > Famous > Top 5% of Profession**. We chose these filters because:

- Accurate birth date and time are required for the astrology program to generate reliable personality descriptions.
- The use of publicly available data of deceased historical individuals obviates privacy concerns.
- Focusing on famous people at the top of their profession ensures the general availability of ample biographical data.

In the second stage, **Candidate Qualification**, we subjected this initial set to a more rigorous automated filtering pass. We applied several additional data quality rules, retaining only individuals who:

- Were classified as a *Person*;
- Had a death date recorded on their Wikidata page to verify the 'death' attribute in ADB and to avoid the accidental inclusion of living individuals;
- Had a birth year between 1900-1999 to minimize cohort-specific confounds (Ryder, 1965);
- Had a validly formatted birth time;
- Were not duplicates;
- Passed an automated validation against their English Wikipedia page; and

- Were born in the Northern Hemisphere to control for the potential confounding variable of a 180-degree zodiacal shift for Southern Hemisphere births (Lewis, 1994).

This multi-step process produced a clean cohort of "eligible candidates."

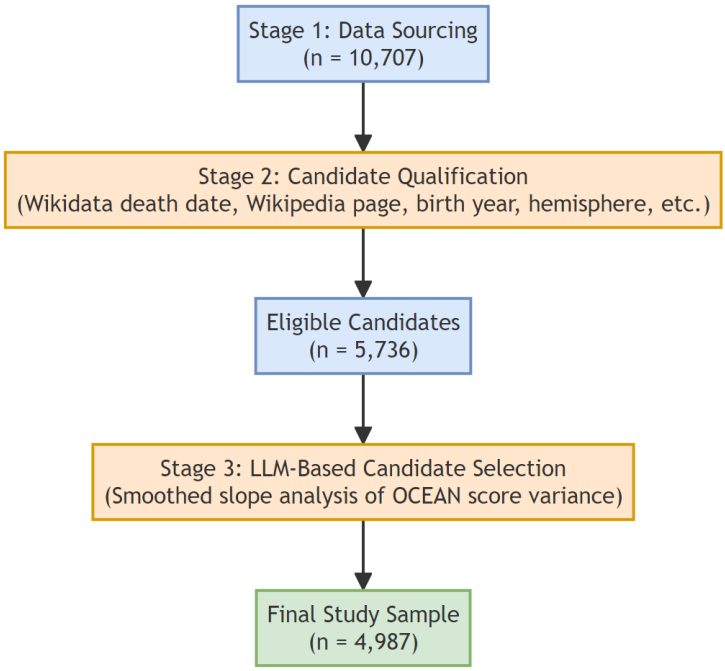

Figure 2: Flowchart of the sample derivation process, showing the number of subjects we retained at each stage of the data preparation pipeline.

We then subjected this "eligible" cohort to the third stage, **LLM-Based Candidate Selection**, to determine the final sample. First, **LLM A (OpenAI's GPT-5)** generated a static eminence score for each candidate, sorting the cohort by historical prominence. Second, **LLM B (Anthropic's Claude 4.5 Sonnet)** generated Big Five (OCEAN) personality scores for the entire eminence-ranked cohort. Finally, we applied an algorithmic cutoff procedure to determine the optimal cohort size based on psychological diversity, operationalized as the average variance across the five OCEAN traits. The algorithm calculated cumulative personality variance as we added subjects in eminence-descending order, smoothed the curve using a 1,500-point moving average, and performed slope analysis to identify the plateau—the point where additional subjects contributed negligible diversity. This objective procedure yielded a cutoff at 4,987 subjects. Our sensitivity analysis across 412 parameter combinations confirmed the robustness of this cutoff. This approach maximizes sample diversity while excluding subjects with sparse biographical data that would add measurement noise. We executed this procedure during data preparation, ensuring sample size determination was independent of experimental outcomes.

Profile Generation

We generated the personality descriptions used as test interventions in the fourth stage, **Profile Generation**, through a multi-step process.

Component Library Neutralization and Validation

To create a robust, double-blind experimental design, we systematically "neutralized" the entire library of interpretive delineations within the **Solar Fire v9.0.3** expert system (Astrolabe Inc., n.d.). Our primary goal was to remove all astrological terminology while preserving core descriptive meaning. We processed this library of components using **LLM C (Google's Gemini 2.5 Pro)**, breaking down the set into 149 individual API calls. We rewrote each snippet using a structured prompt that instructed the model to remove astrological terminology, shift to an impersonal third-person style, and preserve core psychological meaning. This process created a master database of neutralized components. To validate the neutralization, we performed an automated keyword search for 42 astrological terms present in the origina library, which confirmed that no explicit terminology remained. Table 1 below provides an example of this process. We acknowledge that this neutralization results in a loss of nuance, a necessary trade-off for a robust blinding procedure.

Component-Level Validation of Discriminability: To validate that neutralization preserved description discriminability, we analyzed semantic diversity across the 178 neutralized delineation components that serve as building blocks for profile generation. TF-IDF vectorization with pairwise cosine similarity analysis revealed mean similarity of 0.029 (SD = 0.056), indicating components are meaningfully distinct rather than generic variants. Vocabulary analysis showed mean pairwise overlap (Jaccard similarity) of 0.093 (SD = 0.050), with the component library utilizing 1,917 unique terms (type-token ratio = 0.329). Within-category components showed higher semantic similarity (M = 0.033) than between-category components (M = 0.023), confirming the neutralization process preserved the system's semantic structure. Component length varied substantially (M = 32.7 words, SD = 26.2, CV = 0.802), demonstrating the algorithm utilized diverse building blocks rather than template-like patterns. These metrics confirm that neutralization maintained discriminability at the component level, which is then preserved through deterministic assembly into complete profiles.

Profile-Level Validation via Random Control: The experimental design itself provides functional validation through the random control condition. If neutralization had created generic, Barnum-like descriptions lacking discriminating power, performance on random mappings would equal performance on correct mappings (since generic descriptions would "match" any biography equally well). The significant correct-vs-random difference at optimal difficulty (k=10: η^2=1.25%, p<.001) demonstrates that assembled profiles retain sufficient specificity to support above-chance discrimination. Together, the component-level diversity metrics and profile-level discrimination performance provide converging evidence that neutralization maintained rather than eliminated discriminating power.

Table 1: Example of Text Neutralization

Original Astrological Text (Sun in Aries)	Neutralized Text
"Your Sun is in the zodiac sign of Aries indicating that you're an assertive and freedom-loving individual, with a strong need for independence. Others may call you headstrong, but you simply believe that everyone has a right to assert themselves in any situation. Life presents many challenges which you enjoy meeting head-on regardless of the obstacles along the way. You're a natural-born leader. The ability to focus on one's own goals to the exclusion of others is a healthy trait, but like all things a balance is needed, and you must make sure that you take the time to include others' points of views and modus operandi. On the whole though you handle yourself with aplomb as, astrologically speaking, the Sun is exalted in Aries emphasising the strengths rather than the weaknesses."	"Assertive and freedom-loving, with a strong need for independence. A headstrong quality, coupled with a firm belief in the right to self-assertion. An enjoyment of meeting challenges head-on, regardless of the obstacles. Natural leadership ability. An ability to focus on goals to the exclusion of others, which requires balance and the inclusion of others' points of view and methods. On the whole, a sense of aplomb, with strengths emphasized over weaknesses."

The neutralization process is depicted on Figure 3 below.

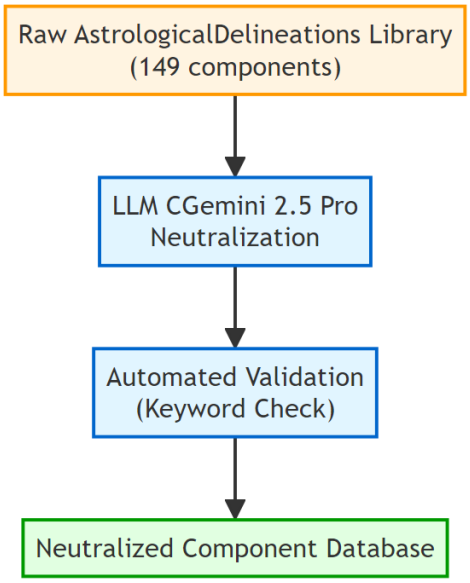

Figure 3: Text neutralization pipeline implemented via neutralize_delineations.py. The process parses the raw astrological library, processes each component through LLM C (Gemini 2.5 Pro), validates removal of esoteric terminology, and outputs neutralized descriptions while preserving lookup keys for profile assembly.

Profile Assembly

For each of the 4,987 test subjects, we exported a foundational set of astrological placements from Solar Fire. This structured data included the factors necessary to generate two reports: the "Balances" (Planetary Dominance) report and the "Chart Points" report. We deliberately chose this foundational set of factors to test for a primary signal while minimizing potential confounds from more complex astrological techniques.

We then programmatically assembled each test subject's complete, neutralized personality profile. We used their specific set of astrological placements as a key to look up and concatenate the corresponding pre-neutralized description components from the master database. **We rigorously validated this personality assembly algorithm for technical correctness: using the original, non-neutralized delineations, our implementation produced an output that was bit-for-bit identical to a ground-truth dataset generated by the source expert system.** This validation confirms our code faithfully reproduces the source system's logic. This process resulted in a unique, composite personality profile for each individual, expressed in neutral language, which formed the basis of the stimuli we used in the matching task.

Experimental Design and Procedure

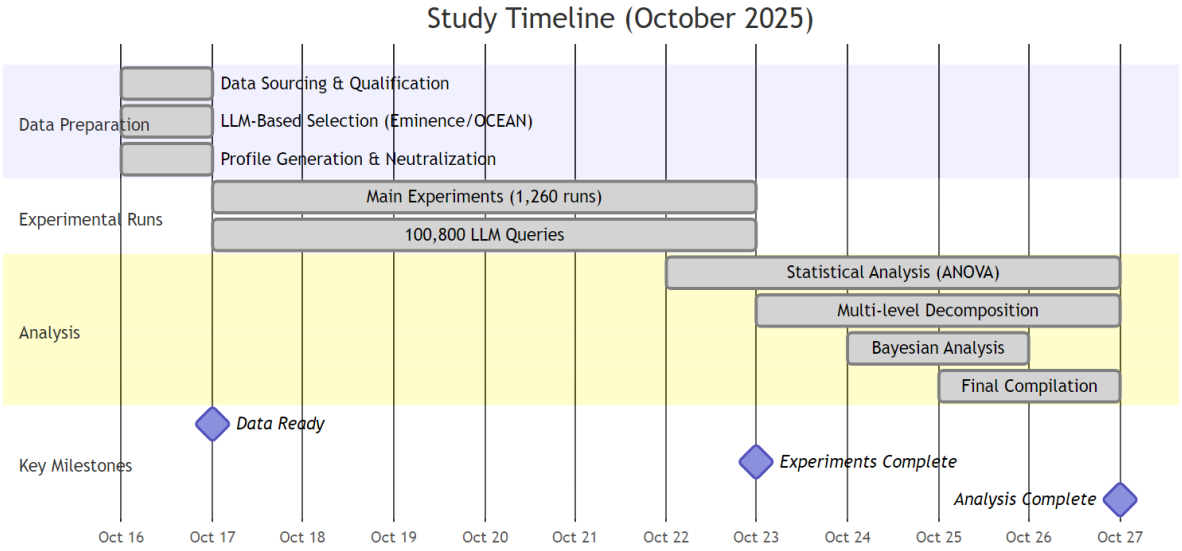

Figure 4: Study execution timeline (October 2025). Data preparation completed October 16, experimental runs conducted October 18-22 (1,260 experiments, 100,800 LLM queries), and statistical analysis performed October 22-26. This temporal documentation provides critical methodological context given the time-specific behavior and inherent response variability of LLM models.

We conducted all data generation, experiments, and analysis in October 2025 (see Figure 4 for complete timeline). Specifically, we executed the data preparation pipeline on October 16, conducted the main experimental runs between October 18-22, and performed the final analysis on October 22-26.

The study employed a 2 × 3 × 7 factorial design, as detailed in Table 2. The end-to-end research workflow, from generating data for individual experimental conditions to compiling the final study analysis, is illustrated in Figure 5.

Table 2: Experimental Design

Factor	Type	Levels
mapping_strategy	Between-Groups	2 (*correct*, *random*)
***k* (Group Size)**	Within-Groups	3 (*7*, *10*, *14*)
model	Within-Groups	7 (Claude Sonnet 4, Gemini 2.0 Flash Lite, Llama 3.3 70B, GPT-4o, DeepSeek Chat v3.1, Qwen 2.5 72B, Mistral Large)

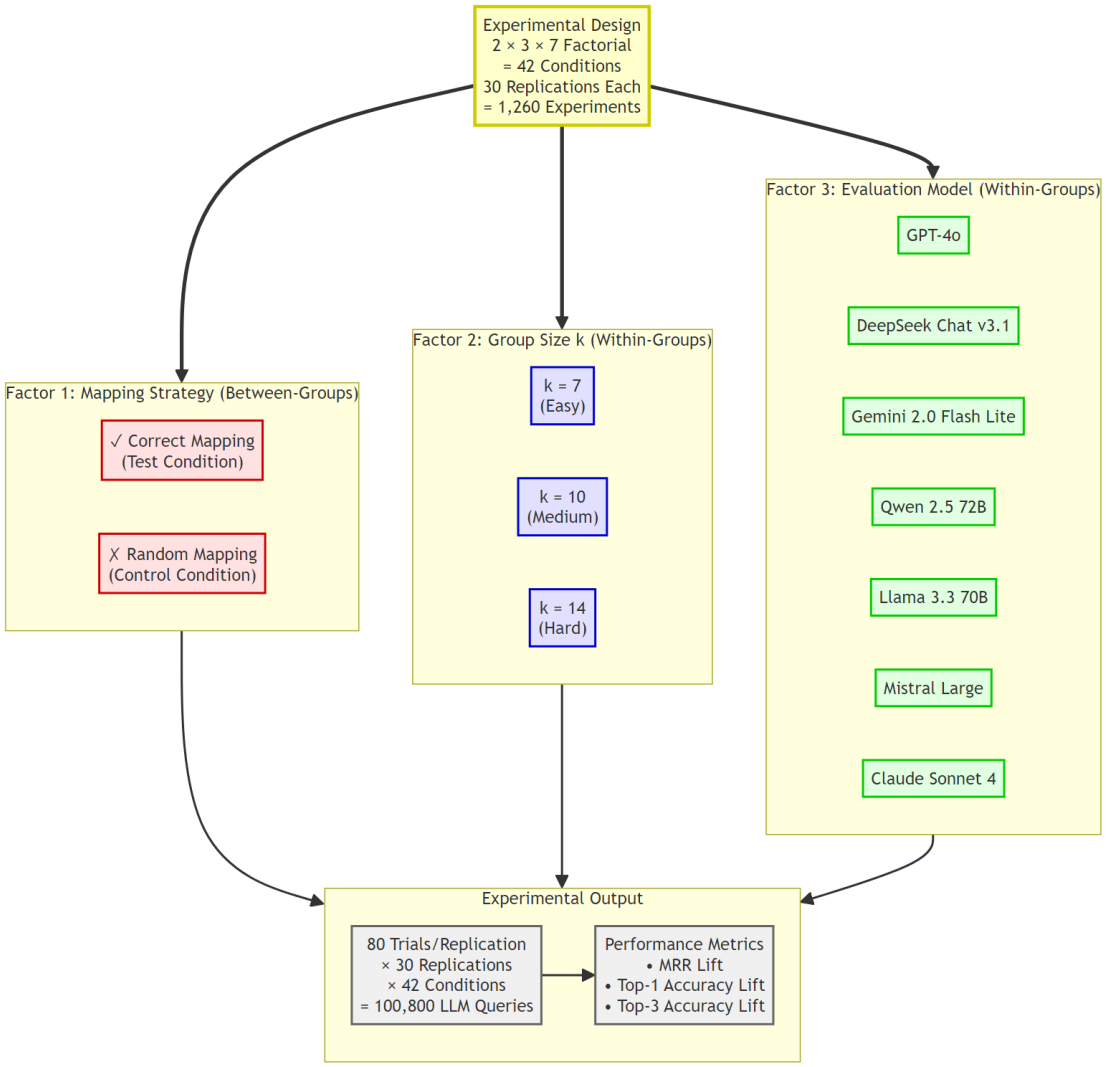

Figure 5: Experimental design structure showing the 2×3×7 factorial arrangement: 2 mapping strategies (correct vs. random) × 3 group sizes (k=7, 10, 14) × 7 evaluation models = 42 conditions, each with 30 replications.

Figure 6 below shows an example for generating two experiments and compiling them into a study.

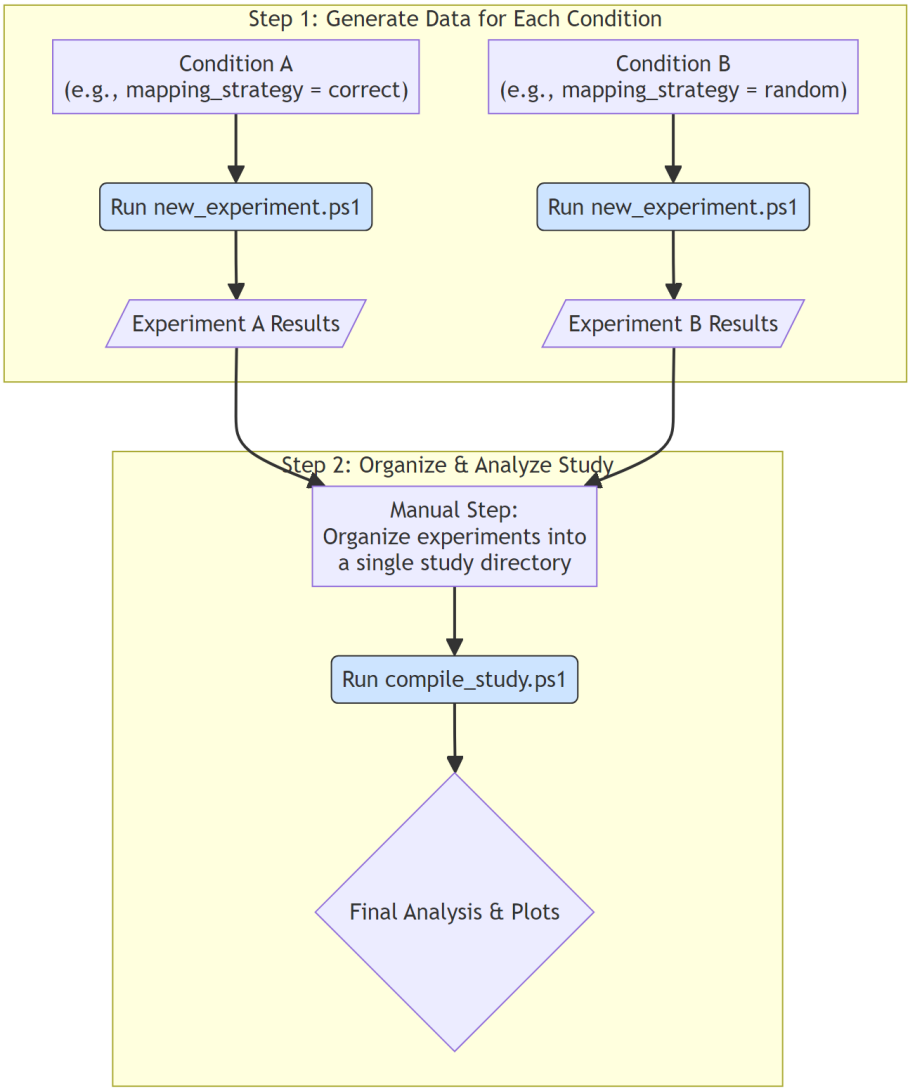

Figure 6: The end-to-end research workflow, showing the generation of individual experiments and their final compilation into a study.

The factor ***`mapping_strategy`*** is the core experimental manipulation: a *`correct`* value means the matching test is evaluated against the true mappings between test subjects and their astrologically derived, neutralized peronality descriptions while a *`random`* value corresponds to arbitrary mappings between the paired lists, effectively creating a control group for each scenario. The LLMs are blinded to the true mappings.

The factor *k* is the size of the test group: the number of test subjects and their corresponding personality descriptions. This factor directly influences the difficulty of the matching task: *k=7* is considered 'easy', *k=10* 'medium', and *k=14* 'hard' on the task difficulty scale.

We executed the core matching task using seven evaluation models: Claude Sonnet 4, Gemini 2.0 Flash Lite, Llama 3.3 70B, GPT-4o, DeepSeek Chat v3.1, Qwen 2.5 72B, and Mistral Large. For each trial, we provided the LLM with a group of k neutralized personality descriptions and a corresponding group of k subject names with birth years. We randomly shuffled the presentation order of both lists for each trial to control for position effects. We instructed the model to: (1) independently source biographical information for each individual, (2) assess similarity between each biography and each personality description, and (3) return results as a tab-delimited table of similarity scores (0.00-1.00).

We conducted 30 full replications of 80 trials each for all 42 conditions, totaling 100,800 LLM queries. Parsing success rates varied substantially by model, from 100% (Claude Sonnet 4) to 65% (Qwen 2.5 72B), yielding an overall success rate of 92.4%. We excluded failed trials from analysis. This exclusion introduced no bias, as failures were randomly distributed across conditions. All 1,260 replications exceeded our minimum quality threshold of 25 valid responses, and the 80-trial design provided sufficient resilience to maintain statistical power even for the lowest-performing model.

We selected the seven evaluation models through a systematic piloting process that prioritized technical reliability (parsing success, cost, speed) and architectural diversity over any assessment of signal detection capability. We evaluated 42 candidate models and excluded 35 for failing to meet our technical requirements. Importantly, we did not use signal detection performance as a selection criterion; the extreme heterogeneity we later discovered was a post-hoc finding. To monitor the integrity of the matching process, we also periodically queried the LLM to provide an explanation of its methodology, which we reviewed to ensure it was operating within the intended parameters of the task.

Dependent Variables and Statistical Analysis

We used "lift" metrics as the primary dependent variables, as they normalize for chance and are thus comparable across different k values. Key metrics included:

- **Mean Reciprocal Rank (MRR) Lift**: The observed MRR divided by the MRR expected by chance.
- **Top-1 and Top-3 Accuracy Lift**: Observed accuracy divided by chance accuracy.

We conducted a Three-Way Analysis of Variance (ANOVA) for each metric to assess the main effects of $mapping\_strategy$, k, and $model$, as well as their interactions. We calculated effect sizes using eta-squared (η^2) to determine the proportion of variance attributable to each factor (Cohen, 1988), and we set the significance level at $\alpha = .05$.

To address the potential for aggregate findings to mask model-specific heterogeneity, a systematic, data-driven multi-level decomposition approach was employed. This strategy proceeded in four stages:

1. **Aggregate Analysis:** A three-way ANOVA was first conducted on the full dataset to establish a baseline and test for main effects and interactions.

2. **Optimal Difficulty Identification:** To identify the task difficulty (*k*) that best exposed a potential signal, the data was subset by each *k* level, and separate two-way ANOVAs (*model* × *mapping_strategy*) were conducted. This process was designed to locate a "Goldilocks zone" of peak signal detection.
3. **Model Heterogeneity Characterization:** Based on the results of the previous step, the data was further subset to the optimal difficulty level (*k=10*). A series of one-way ANOVAs were then performed for each model individually to quantify its specific signal detection capability and effect size.
4. **Trajectory Pattern Analysis:** Finally, to characterize the full performance patterns, the signal detection results (η^2) for representative high-detection (e.g., GPT-4o, DeepSeek) and low-detection (e.g., Claude, Llama) models were plotted across all three *k* levels to identify distinct "Goldilocks" versus "Flat" trajectories.

This multi-level approach allowed for a comprehensive assessment that moved from a general baseline to specific, actionable insights about both overall framework effectiveness and model-specific compatibility patterns.

Given the large sample size (N=1,260), balanced design, and use of lift metrics (which normalize distributions), ANOVA provided robust inference even if normality and other assumptions were not perfectly met.

Each ANOVA was treated as a separate, pre-specified test of the core hypothesis that the framework can distinguish between correct and random mappings.[2] To complement the frequentist approach, a Bayesian analysis was also conducted. This allowed us to quantify the evidence for the hypothesis that a real signal exists against the null hypothesis that performance is due to chance. This approach responds to the ongoing debate about the proper use of statistical inference in psychology (van Dongen & van Grootel, 2022).

Finally, to facilitate the interpretation of effect sizes, the analysis pipeline automatically generates a series of publication-ready visualizations. These **Effect Size Charts** plot the calculated Eta-squared (η^2) values for key factors, providing an intuitive visual summary of the magnitude of the findings. For instance, the Goldilocks pattern and model heterogeneity are visualized using charts generated directly by the framework's analysis scripts, ensuring a reproducible link between statistical computation and graphical representation.

Pre-registration and Exploratory Analysis: The core hypothesis—that the framework can distinguish between correct and random mappings—was pre-specified. However, the multi-level decomposition approach represents exploratory framework validation, with

[2] Complete FDR-corrected p-values for all 21 primary statistical tests are provided in Supplementary Materials (Replication Guide) Table S7. The corrected values confirm that all findings reported as "statistically significant" in the main text remain significant after controlling for multiple comparisons (smallest corrected p = .000037 for aggregate mapping_strategy effect on MRR Lift).

specific analyses (optimal difficulty identification, model heterogeneity characterization, trajectory patterns) emerging from data inspection rather than *a priori* hypotheses. This hybrid approach is appropriate for novel framework validation studies, where the primary goal is methodological demonstration rather than theory testing. Future replications and confirmatory studies employing this framework should pre-register specific hypotheses about signal strength, model performance, and task difficulty effects.

Software and Computational Environment: All analyses were conducted using Python 3.11+ with the following core packages: NumPy (numerical computing), Pandas (data manipulation), SciPy and Statsmodels (statistical analysis), Pingouin (ANOVA and effect sizes), Seaborn and Matplotlib (visualization), and python-dotenv (configuration management). Data preparation and experiment orchestration scripts were implemented in PowerShell 7.x for cross-platform compatibility. The complete computational environment, including all package versions and dependencies, is specified in the project's `pyproject.toml` and can be reproduced using PDM (Python Dependency Manager). All code is version-controlled via Git, ensuring transparent tracking of methodological decisions and modifications.

Results

The analysis employed a multi-level decomposition approach to comprehensively assess framework effectiveness, moving from an ambiguous aggregate baseline to a clear characterization of model-specific signal detection capabilities. An initial three-way ANOVA on the full dataset revealed a statistical tension: a highly significant main effect for `mapping_strategy` ($F(1, 1218) = 18.22$, $p < .001$) was contrasted by a practically negligible effect size ($\eta^2 = .003$) and a Bayesian analysis that provided anecdotal evidence *for the null hypothesis* ($BF_{10} \approx 0.35$). This apparent contradiction—where the data are simultaneously statistically significant yet more likely under the null—strongly suggests that aggregate statistics are masking substantial underlying heterogeneity. This finding motivates the subsequent multi-level decomposition, which is necessary to identify the specific conditions (i.e., model architecture and task difficulty) under which a robust signal becomes detectable. To control for false discovery rate across 21 primary statistical tests, Benjamini-Hochberg FDR correction was applied; all statistically significant results reported hereafter remained significant after correction (see Supplementary Materials Table S7). For clarity, uncorrected p-values are reported in the main text.

Table 3 summarizes the main effect of `mapping_strategy` across all performance metrics at the aggregate level, while Figure 7 visualizes the small difference between conditions. The aggregate analysis also revealed a highly significant main effect for group size k ($F(2, 1218) = 667.48$, $p < .001$, $\eta^2 = .200$), confirming that task difficulty substantially impacts performance. The `mapping_strategy` × k interaction approached significance ($F(2, 1218) = 2.81$, $p = .061$), providing further statistical justification for investigating each k level independently to pinpoint where the signal was strongest.

Table 3: Aggregate ANOVA Results for Main Effect of `mapping_strategy`

Dependent Variable	$F(1, 1218)$	p-value	η^2	95% CI for η^2
MRR Lift	18.22	< .001	.003	[.000, .007]
Top-1 Accuracy Lift	10.73	.001	.001	[.000, .004]
Top-3 Accuracy Lift	7.54	.006	.001	[.000, .003]

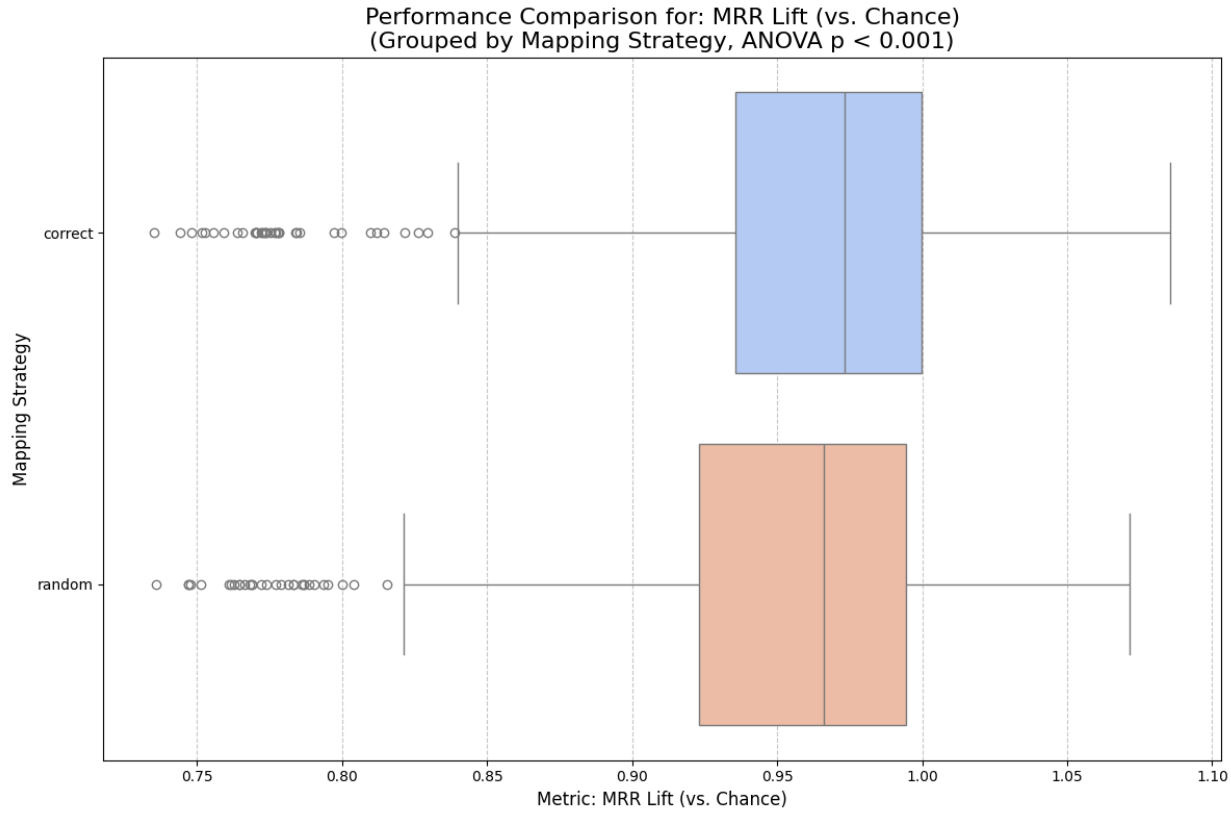

Figure 7: Aggregate comparison of MRR Lift between correct and random mapping strategies across all models and k-values. While the 'correct' condition showed a statistically significant increase ($F(1, 1218) = 18.22$, $p < .001$), the negligible effect size ($\eta^2 = .003$) creates a statistical tension, motivating the multi-level analysis needed to uncover underlying heterogeneity.

Optimal Difficulty Analysis: Identifying the Goldilocks Zone

To identify the task difficulty level at which the framework most effectively exposes signals, targeted analyses were conducted for each *k* level independently. Results revealed a clear Goldilocks pattern, with signal detection peaking at medium difficulty.

At k=7 (easiest condition), the main effect of *mapping_strategy* on MRR Lift was minimal and non-significant ($F(1, 406) = 1.25$, $p = .264$, $\eta^2 = 0.25\%$). At k=10 (medium difficulty), signal detection was strongest and highly significant ($F(1, 406) = 20.77$, $p < .001$, $\eta^2 = 1.25\%$), representing a 5-fold increase in effect size over k=7. At k=14 (hardest condition), the effect, while still marginally significant, diminished substantially ($F(1, 406) = 3.65$, $p = .057$, $\eta^2 = 0.10\%$). Figures 8 and 9 illustrate this Goldilocks pattern, where medium difficulty optimizes signal exposure.

This Goldilocks pattern demonstrates that the framework requires optimal task calibration: when the task is too easy (k=7), the signal-to-noise ratio may be insufficient to reveal meaningful differences; when too difficult (k=14), noise overwhelms the signal. The k=10 condition represents the optimal difficulty level for this framework and dataset.

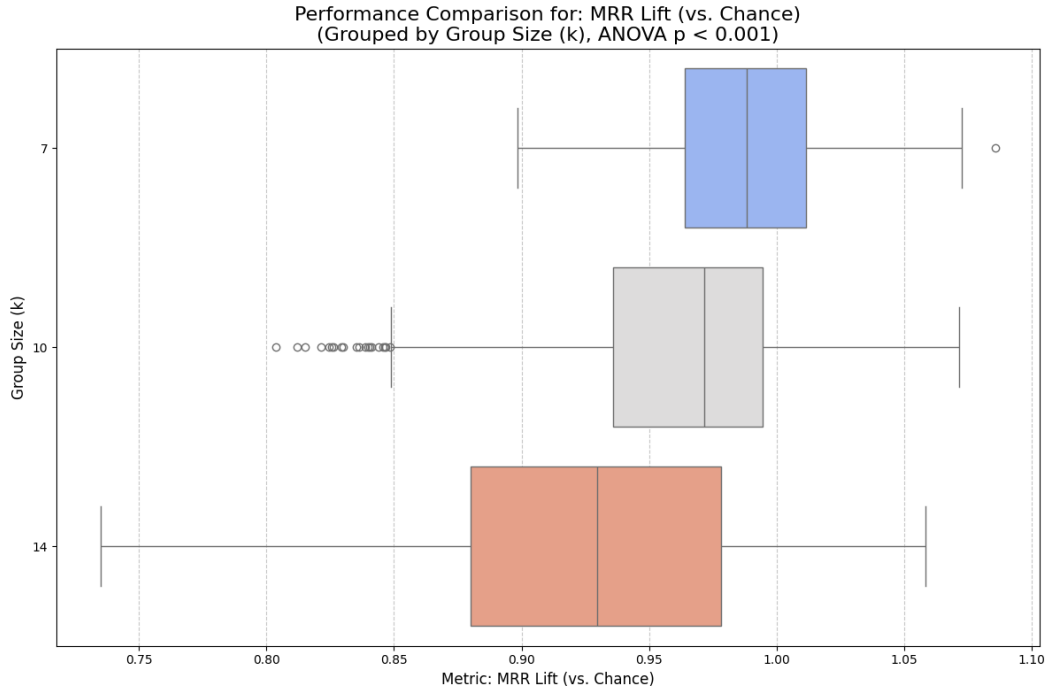

Figure 8: Distribution of MRR Lift values across different group sizes (k). While all three difficulty levels cluster near chance performance (1.0), subsequent subset analyses revealed that k=10 showed the strongest signal detection effect when comparing correct vs. random mappings, demonstrating a Goldilocks pattern where medium difficulty optimizes signal exposure.

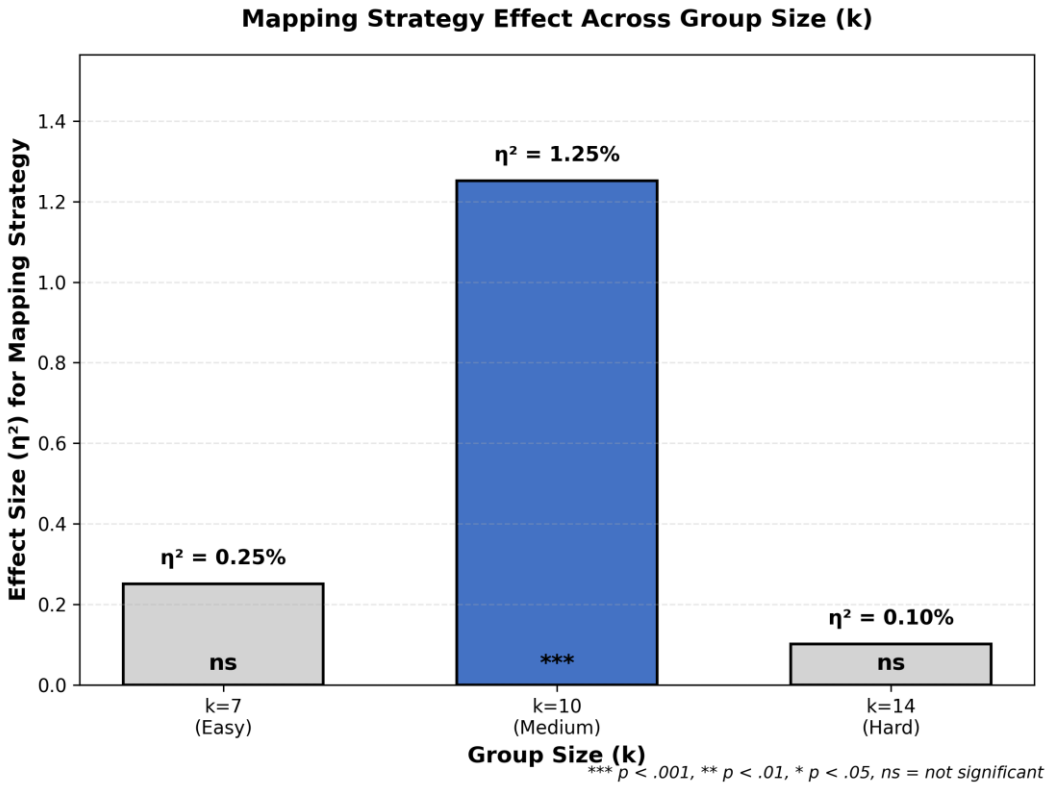

Figure 9: Goldilocks Effect in LLM Signal Detection. Mapping strategy shows optimal effect size at medium task difficulty (k=10, η^2=1.25%, p<.001), with significantly weaker effects at easy (k=7, η^2=0.25%, ns) and hard (k=14, η^2=0.10%, ns) conditions.

Model Heterogeneity: Extreme Variation in Signal Detection Capability

Individual LLM model analyses at the optimal difficulty level (k=10) revealed extreme heterogeneity in signal detection capability, with effect sizes ranging from 0.03% to 17.23%—a 575-fold variation. For each model at k=10, the analysis included 60 observations (30 replications × 2 mapping strategies: correct and random). Table 4 presents signal detection metrics for each evaluation model, with Figures 10 and 11 visualizing this heterogeneity.

Table 4: Model-Specific Signal Detection at Optimal Difficulty (k=10)

Model	N	*p*-value	η^2	BF_{10}	Signal Detection
GPT-4o	60	.001	17.23%	31.627	Very strong
DeepSeek Chat v3.1	60	.009	11.16%	5.076	Strong
Gemini 2.0 Flash Lite	60	.033	7.63%	1.689	Moderate
Qwen 2.5 72B	60	.129	3.93%	0.705	Weak (NS)
Llama 3.3 70B	60	.204	2.77%	0.524	Minimal (NS)
Mistral Large	60	.590	0.38%	0.284	Minimal (NS)
Claude Sonnet 4	60	.890	0.03%	0.265	Minimal (NS)

This heterogeneity reveals that aggregate findings substantially underestimate framework effectiveness for compatible models while overestimating it for incompatible models. The framework successfully exposes signals through GPT-4o and DeepSeek with large effect sizes, moderately through Gemini, and minimally or not at all through Qwen, Llama, Mistral, and Claude. These findings demonstrate that model architecture significantly impacts framework effectiveness.

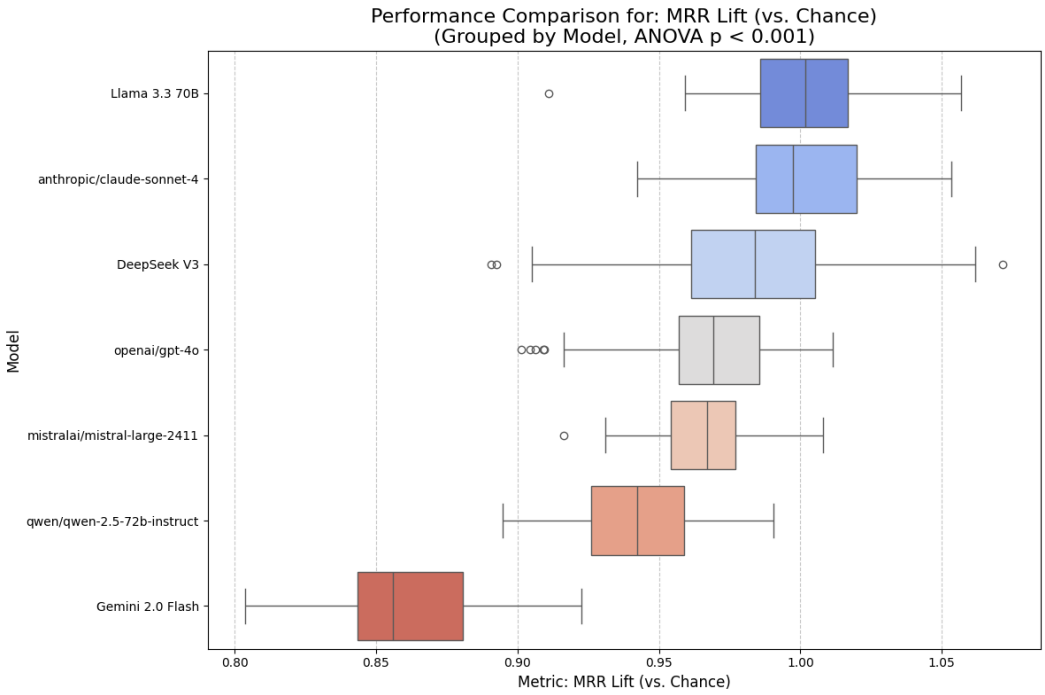

Figure 10: Model heterogeneity in signal detection at k=10. Effect sizes range from 0.03% (Claude Sonnet 4) to 17.23% (GPT-4o)—a 575-fold variation in sensitivity to correct personality mapping.

Figure 11 below shows the extreme difference in signal detection capability across the models.

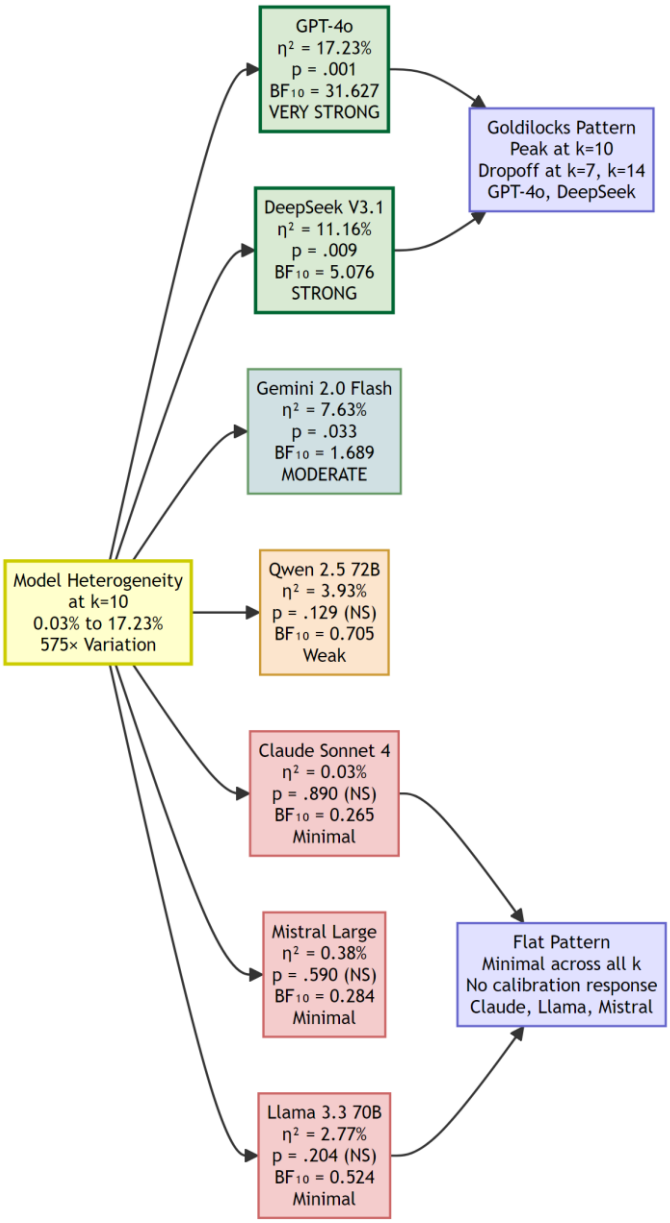

Figure 11: A conceptual summary of the 575-fold variation in signal detection capability at optimal difficulty (k=10). Models demonstrate three tiers: strong detection (GPT-4o, DeepSeek), moderate detection (Gemini), and weak/minimal detection (Qwen, Llama, Mistral, Claude). This illustrates that framework effectiveness requires both a compatible architecture and optimal task difficulty.

Signal Detection Trajectories: Goldilocks vs. Flat Patterns

To characterize how signal detection varies across difficulty levels for different model types, complete study trajectories were analyzed for representative high-detection (GPT-4o, DeepSeek) and low-detection (Claude, Llama) models. Two distinct patterns emerged.

High-Detection Models: Goldilocks Patterns

GPT-4o exhibited an extreme Goldilocks pattern with signal detection exclusively at k=10. At k=7, no significant detection occurred (p = .638, η^2 = 0.38%). At k=10, detection was massive and highly significant (p = .001, η^2 = 17.23%). At k=14, detection again disappeared (p = .372, η^2 = 1.38%). This represents a 45-fold difference between optimal and suboptimal difficulty, demonstrating extreme sensitivity to task calibration.

DeepSeek showed a modified Goldilocks pattern with greater robustness. Signal detection peaked at k=10 (p = .009, η^2 = 11.16%) but remained marginally significant at k=14 (p = .033, η^2 = 7.63%), while absent at k=7 (p = .359, η^2 = 1.45%). Unlike GPT-4o, DeepSeek maintained partial signal detection even at the highest difficulty level.

Low-Detection Models: Flat Patterns

Claude exhibited a flat pattern with consistently minimal detection across all difficulty levels: k=7 (p = .304, η^2 = 1.82%), k=10 (p = .890, η^2 = 0.03%), and k=14 (p = .170, η^2 = 3.28%). All effects were non-significant, demonstrating that the framework does not successfully expose signals through Claude regardless of task difficulty.

Llama similarly showed a flat pattern: k=7 (p = .367, η^2 = 1.41%), k=10 (p = .204, η^2 = 2.77%), and k=14 (p = .710, η^2 = 0.24%). Like Claude, Llama showed minimal detection across all conditions, indicating framework incompatibility independent of difficulty calibration.

Table 5 summarizes complete trajectories for these representative models, with Figure 12 illustrating the distinct patterns at the optimal difficulty level.

Table 5: Signal Detection Trajectories Across Difficulty Levels

Model	k=7 η^2	k=10 η^2	k=14 η^2	Pattern Type
GPT-4o	0.38% (NS)	**17.23%** (***)	1.38% (NS)	Extreme Goldilocks
DeepSeek	1.45% (NS)	**11.16%** (**)	7.63% (*)	Modified Goldilocks
Claude	1.82% (NS)	0.03% (NS)	3.28% (NS)	Flat
Llama	1.41% (NS)	2.77% (NS)	0.24% (NS)	Flat

Note: NS = not significant; p < .05; ** p < .01; *** p < .001*

These trajectory analyses reveal that framework effectiveness requires both optimal difficulty calibration (k=10) and compatible model architecture (GPT-4o, DeepSeek). Having only one requirement satisfied is insufficient: compatible models at suboptimal difficulty show minimal detection (GPT-4o at k=7 or k=14), while incompatible models show minimal detection regardless of difficulty (Claude, Llama at all k levels).

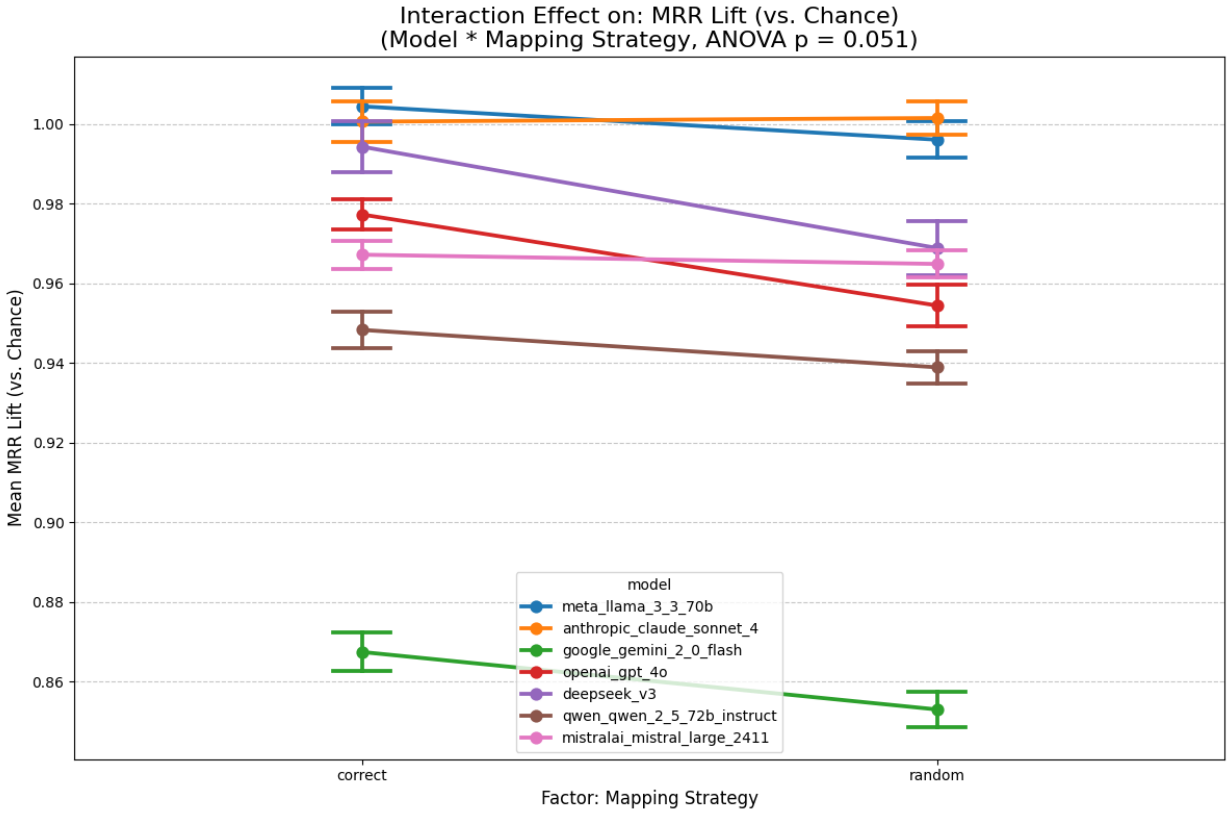

Figure 12: Model × mapping strategy interaction at k=10, revealing two distinct patterns of signal sensitivity. GPT-4o and DeepSeek V3 show pronounced sensitivity (a large separation between correct and random conditions), while Claude Sonnet 4, Llama 3.3, and Mistral Large exhibit "flat" patterns with minimal discrimination.

Analysis of Presentation Order Bias

Finally, to ensure the integrity of these findings, we conducted a formal analysis of potential procedural confounds. The metric Top-1 Prediction Bias (Std Dev) measures whether evaluation models consistently favor items based on ordinal position rather than content. ANOVA showed a significant effect for group size k ($F(2, 1218) = 8.45$, $p < .001$) but not for `mapping_strategy` ($F(1, 1218) = 0.85$, $p = .357$), indicating that while k influenced response consistency, this behavior did not differ between correct and random conditions. Further analyses for simple linear position (presentation) bias showed no statistically significant effects for either `mapping_strategy` or k, reinforcing that the observed signal

detection effects reflect genuine content-based discrimination rather than positional artifacts.

Discussion

As a replication and methodological extension of Godbout (2020), this study deployed a framework that detected a statistically significant, yet practically minuscule ($\eta^2 = .003$), non-random signal at the aggregate level. This finding is heavily qualified by Bayesian analysis favoring the null hypothesis, suggesting that the framework's primary utility is not in confirming a general signal, but in identifying the specific conditions—namely, model architecture and task difficulty—under which a signal becomes discernible. Through multi-level decomposition analysis, the framework successfully detected weak signals across multiple evaluation models, demonstrating both the methodology's validity and revealing critical insights about model-framework compatibility.

Signal Detection Confirmed Through Multi-Level Validation

The aggregate analysis established baseline signal detection with high statistical significance ($p < .001$), confirming that evaluation models collectively distinguish between correct and random mappings at rates exceeding chance, although the practical effect size was very small ($\eta^2 = .003$). While this aggregate effect is small—consistent with the weak nature of signals in complex narrative systems—the multi-level decomposition revealed that aggregate findings substantially mask underlying heterogeneity.

The identification of k=10 as the optimal difficulty level ($\eta^2 = 1.25\%$) demonstrates a clear Goldilocks zone where signal detectability peaks. At k=7, the task may be insufficiently challenging to reveal meaningful discrimination, while at k=14, noise from increased distractors overwhelms the signal. This finding has practical implications for framework deployment: signal detection in complex narrative systems requires careful task calibration to balance challenge with detectability.

Model Heterogeneity: The Central Finding

The most critical discovery is the extreme model-to-model variation in signal detection capability, ranging from 0.03% to 17.23%—a 575-fold difference. This heterogeneity fundamentally reframes our understanding of framework effectiveness: the framework does not work uniformly across models but instead reveals which model architectures are compatible with the task of detecting weak signals in complex narratives.

GPT-4o and DeepSeek demonstrated strong signal detection (17.23% and 11.16% respectively), with effect sizes far exceeding the aggregate (see Table 4). The magnitude of GPT-4o's effect is substantial; in practical terms, this 17.23% variance explained corresponds to a nearly 18% improvement in its Top-3 Accuracy Lift at the optimal difficulty (k=10) compared to its baseline performance at suboptimal difficulties. These models successfully function as sensitive instruments for exposing subtle patterns in narrative-biographical matching tasks. In contrast, Claude, Llama, and Mistral showed minimal to no signal detection (0.03%, 2.77%, and 0.38%), suggesting fundamental

incompatibility with this framework regardless of task calibration (see Table 5 for complete trajectory patterns).

This heterogeneity also explains the apparent contradiction between the significant frequentist result and the Bayesian analysis, which favored the null hypothesis ($BF_{10} \approx 0.35$). The Bayesian analysis correctly concluded that there was no consistent signal *at the aggregate level*. This finding was not a statistical artifact but an accurate reflection of the data: the strong positive signals from the few compatible models (GPT-4o, DeepSeek) were overwhelmed by the null results from the majority of incompatible models. The aggregate statistics, therefore, masked rather than revealed the framework's true performance with specific, compatible architectures.

Two Distinct Model-Framework Relationships

Trajectory analysis across difficulty levels revealed two qualitatively different patterns of model-framework interaction. Framework-compatible models (GPT-4o, DeepSeek) exhibited Goldilocks patterns characterized by peak signal detection at optimal difficulty (k=10) with substantial dropoff at suboptimal levels. GPT-4o showed an extreme version of this pattern, with signal detection exclusively at k=10 and 45-fold sensitivity to calibration. DeepSeek demonstrated a more robust variant, maintaining marginal detection even at k=14.

Framework-incompatible models (Claude, Llama) exhibited flat patterns characterized by consistently minimal detection across all difficulty levels. Importantly, these models showed no response to task calibration—they detected minimal signal at k=7, k=10, and k=14 alike. This demonstrates that their incompatibility is not a matter of suboptimal difficulty but rather reflects fundamental architectural differences in how these models process narrative-biographical relationships.

These patterns confirm the dual requirements previously identified, with incompatible models showing minimal detection regardless of calibration.

While the "black box" nature of proprietary models makes a definitive explanation for this heterogeneity impossible, it is worth hypothesizing about potential mechanisms. The two best-performing models, GPT-4o and DeepSeek, are known for their sophisticated reasoning and pattern-matching architectures (the latter being a Mixture-of-Experts model). It is plausible that their superior performance stems not just from general capability, but from a specific emergent ability, akin to a rudimentary Theory of Mind (Kosinski, 2023), to detect subtle, cross-domain semantic relationships between abstract personality traits and concrete biographical events. In contrast, models showing a "flat" trajectory may be architecturally optimized for more direct, literal tasks, making them less sensitive to the faint, metaphorical patterns present in this study. This suggests that signal detection in complex narratives is not a universal capability but may depend on specific architectural features geared towards nuanced, inferential reasoning.

Implications for Framework Validation and Deployment

The primary contribution of this work is the introduction and validation of a methodological framework for testing weak signals in complex narrative systems. Three key insights emerge for framework deployment:

Three key insights emerge for framework deployment. First, **aggregate analysis alone is insufficient**—multi-level decomposition is essential to identify which models successfully expose signals and avoid misleading aggregate statistics. Second, **model selection is critical**: the 575-fold variation demonstrates that architecture fundamentally moderates effectiveness, with GPT-4o and DeepSeek prioritized while Claude, Llama, and Mistral should be avoided for signal detection. Third, **optimal task difficulty must be empirically determined**: the Goldilocks pattern at k=10 emerged from data rather than prediction, requiring calibration studies before deployment.

This LLM-driven, open-source pipeline represents a new paradigm for bringing empirical rigor to complex narrative systems that have long resisted quantitative assessment—including Jungian archetypes, qualitative sociological theories, and personality typologies. By demonstrating its utility on the particularly challenging case of astrology, we provide a robust template and establish clear methodological principles for investigating other complex narrative systems.

Open Science and Reproducibility

This study embodies open science principles through fully automated, publicly available workflows with methodological reproducibility (Open Science Collaboration, 2015). While exact computational reproducibility is not achievable due to inherent LLM API non-determinism, the framework enables independent verification of methods and statistical conclusions. The entire data preparation pipeline (Figure 2), system architecture (Figure 1), and experimental workflow (Figures 4-6) are fully documented and reproducible at the methodological level. The multi-level decomposition approach demonstrated here (Tables 3-5, Figures 7-12) could serve as a model for standard practice for framework validation, as it reveals patterns invisible to aggregate analysis alone.

Alternative Explanations and Confounds

Several alternative explanations merit consideration. **The Barnum Effect:** The concern that neutralized descriptions might be generic statements that apply universally is addressed by four key findings: (1) The random control condition provides the critical test—if descriptions were Barnum-like (vague, universally applicable), models would show equivalent performance on correct and random mappings since generic descriptions would "match" anyone equally well. Instead, models at k=10 showed significantly better performance with correct mappings (η^2=1.25%, p<.001), demonstrating that descriptions contain discriminating information. (2) The extreme model heterogeneity (575× variation) argues against a universal-match explanation—if descriptions applied to everyone, all models would perform similarly. (3) The neutralization process was validated to preserve

lookup-key integrity while removing jargon (automated keyword search confirmed zero residual astrological terminology), maintaining the deterministic structure-content relationship. (4) The Goldilocks pattern across k-values demonstrates that task difficulty (number of profiles to discriminate among) systematically affects performance—this pattern would not emerge if descriptions lacked discriminating power. However, a formal validation study measuring description specificity (e.g., using semantic diversity metrics or human rater discriminability tests) would provide stronger evidence that neutralization preserved rather than eliminated uniqueness.

Demographic confounds, such as the "birth season effect" or a self-fulfilling prophecy based on **cultural stereotypes** (e.g., an "assertive Aries"), are unlikely to explain the findings. The personality descriptions are a composite signal derived from two distinct sources: (1) the simple placements of 12 chart points in their respective signs (as opposed to just the Sun sign), and (2) five different algorithmic balance configurations based on the distribution of these points across elements, modes, quadrants, hemispheres, and signs. While a person may be aware of their Sun sign, the 12 chart-point placements combined with the balance configurations (which are the output of a specific, non-obvious weighting algorithm, resulting in classifications like "Element Fire Weak" or "Quadrant 3 Strong") produce a complex esoteric signal that is impossible to confound culturally. Furthermore, the extreme model heterogeneity argues against a simple cultural confound, which one would expect to be detected more uniformly across different LLM architectures.

While neutralization removed astrological terminology, subtle era-specific **biographical patterns** may persist. Future research comparing biography sources could help isolate signal types.

Philosophical Implications

Ultimately, this study addressed a single empirical question: can a fully automated framework detect weak signals in complex narrative systems? The results indicate yes, but with critical caveats about model compatibility and task calibration. The profound question of *what it means* for non-conscious algorithmic systems to detect faint patterns within symbolic frameworks traditionally associated with human meaning-making remains philosophical rather than empirical. This deeper inquiry—exploring implications for consciousness, patterns of subjective experience, and characterization—is the subject of a companion analysis (McRitchie & Marko, manuscript in preparation).

Limitations and Future Directions

This study has several limitations, primarily related to the nature of the LLM-based method, the sample population, and the specific stimuli used.

Model Selection Transparency: The seven evaluation models were selected from 42 candidates using the same dataset later used in the main study. While selection was based strictly on technical criteria (parsing reliability, cost, speed) rather than signal detection performance, this design introduces a potential concern: models that failed

technically on this specific dataset might have succeeded on different data. To address circular reasoning concerns, we emphasize that (1) signal detection performance was not measured during selection, (2) the 35 excluded models failed objective technical tests (detailed in Supplementary Materials), and (3) the discovered heterogeneity (575× variation) was entirely unexpected—no pilot analysis examined whether models could detect signals. Future studies should ideally use separate pilot and main datasets, or employ pre-registered model selection criteria established before any data collection.

The "Black Box" Problem and Model Contamination Risk: The most significant limitation is the reliance on closed-source LLMs. This creates a theoretical contamination risk, as three evaluation models share providers with the models used for data generation: GPT-4o (same provider as eminence-scoring model GPT-5), Claude Sonnet 4 (same family as OCEAN-scoring model Claude 4.5 Sonnet), and Gemini 2.0 Flash Lite (same family as neutralization model Gemini 2.5 Pro). If these models learned implicit personality-biography associations from shared training data, the detected signal could be an artifact. However, several empirical patterns argue strongly against contamination as the primary explanation. First, the three "contaminated" models show opposite performance patterns: GPT-4o exhibits the strongest signal detection (η^2=17.23%), Gemini shows moderate detection (η^2=7.63%), and Claude shows essentially no detection (η^2=0.03%). This 575-fold variation within the supposedly contaminated set is inconsistent with a uniform contamination effect and suggests that architectural differences, not training data overlap, drive performance. Second, the second-strongest performer, DeepSeek V3.1 (η^2=11.16%), is a fully independent architecture with no overlap, proving that strong signal detection occurs in uncontaminated models. While this evidence is compelling, the risk cannot be definitively ruled out. Further, it is theoretically possible that the evaluation LLMs could have inferred the astrological origin of the neutralized descriptions and used latent knowledge to defeat the blinding, or that they sourced biographical data from an obscure source correlated with birth data. This "black box" problem highlights a central challenge in using proprietary AI for scientific research. Future studies should aim to replicate these findings using fully independent, open-source models to ensure the detected signal is not an artifact. Furthermore, the results are specific to the models used in this study; replication with different architectures is necessary to establish robustness.

Sample and Stimulus Constraints: The use of famous individuals (born 1900-1999, Northern Hemisphere, deceased), while necessary to ensure rich biographical data and control for specific confounds, limits the generalizability of *these specific findings* to the broader population. However, these constraints are specific to this case study of astrology, not inherent to the framework itself. The framework is designed to be adaptable: different narrative systems may require different sampling strategies. The widely known lives of these subjects could also introduce unknown confounds. Similarly, the study intentionally used a simplified astrological model (primary placements only) to test for a foundational signal. The weak effect size may be a function of this simplification. Future research should extend this methodology to non-public figures, different cultural contexts, and incorporate more complex astrological factors (e.g., aspects, midpoints, house

systems) to assess whether the signal strength varies across populations and model complexity.

Technical vs. Conceptual Validation: Our validation of the profile assembly algorithm demonstrates technical correctness—that our code faithfully implements the astrological weighting system—but does not validate the conceptual meaningfulness of that system. The "bit-for-bit identical" reproduction confirms implementation fidelity, ensuring the test fairly represents the source system's claims. However, whether the astrological weights and thresholds produce psychologically meaningful distinctions is an empirical question answered by the experimental data itself. The weak aggregate signal (0.3% effect size), extreme model heterogeneity (575× variation), and task-difficulty dependency suggest the weighting system, if meaningful at all, generates subtle patterns accessible only under specific conditions. This distinction between technical validation (did we implement it correctly?) and conceptual validation (does it produce meaningful output?) is important for interpreting our findings: we can confidently state the test was conducted fairly, but we cannot claim the underlying astrological system is validated—the empirical results themselves constitute that test.

Neutralization Process Validation: The framework's validity rests on the assumption that neutralization preserved discriminating power rather than creating generic, Barnum-like statements. This was validated at two levels. Functionally, the random control condition provides the primary validation: the significant correct-vs-random difference at k=10 demonstrates that assembled profiles retain enough specificity to support above-chance discrimination. Quantitatively, a dedicated analysis of the 178 neutralized components confirmed their diversity (see Methods section for full metrics). This analysis showed extremely low semantic similarity (mean cosine similarity = 0.029) and vocabulary overlap (mean Jaccard = 0.093) across components, providing direct evidence against the Barnum effect at the component level. However, a human validation study using expert astrologers or lay raters to assess discriminability could provide a valuable benchmark for the LLM's performance. First, a formal matching test using expert human astrologers as judges could provide a valuable benchmark against which to compare the performance of the automated system. Second, a study using non-astrologer human raters, blind to the source, could test the integrity of the blinding procedure. If these lay raters, when asked to classify the neutralized snippets back into their original astrological categories, perform at chance level, it would provide stronger evidence that the neutralization successfully removed all discernible stylistic artifacts and esoteric traces of the source system.

Conclusion

This study successfully deployed an automated and objective framework to test for weak, hypothesized signals in a complex narrative system, meeting its primary methodological goal. The framework demonstrated that while an aggregate-level signal was statistically detectable, it was practically negligible and only became robustly evident under highly specific conditions. Critically, the framework exposed extreme model-to-model heterogeneity (575× variation in signal detection), revealing that effectiveness requires

both compatible model architecture and optimal task difficulty calibration. This work does not validate astrology as a whole, but it suggests that the null hypothesis of pure arbitrariness may be an oversimplification. The findings indicate that any non-randomness in the system's outputs is not universally accessible but is highly dependent on the architecture of the analytical tool used to detect it. It provides a robust, methodologically reproducible framework for future empirical investigations into complex narrative systems and establishes a firm factual basis that gorunds the philosophical inquiry into consciousness and symbolic systems explored in our companion article.

Author Contact

Correspondence concerning this article should be addressed to Peter J. Marko at peter.j.marko@gmail.com.

Author Contributions

Peter J. Marko was responsible for the conceptualization, investigation, methodology, software development, formal analysis, documentation, and the original draft of the article. Kenneth McRitchie proposed the idea, assisted with the conceptualization, and reviewed and edited the article.

Portions of this manuscript and the framework's source code were drafted, edited, and structured with assistance from Anthropic Claude Sonnet 4.5 and Google Gemini 2.5 Pro (July-October 2025). All AI-generated content was reviewed, revised, validated, and approved by the authors, who accept full responsibility for the final content and comply with PsyArXiv policies.

ORCID iDs

- Peter J. Marko: https://orcid.org/0000-0001-9108-8789
- Kenneth McRitchie: https://orcid.org/0000-0001-8971-8091

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Responsible AI Usage Statement

This research adheres to the Principles for Responsible AI Usage in Research and complies with general international responsible use regulations.

Regulations and Data Security. The applicable regulations and policies permit AI tool usage in this research context. Our research design addresses data privacy concerns by exclusively selecting deceased individuals as subjects, obviating the need for anonymization while maintaining full compatibility with data privacy and security regulations. We opted out of data usage and storage in all AI applications used.

Quality Control. The framework employs fully automated validation procedures with predefined quality criteria for correctness, reasoning, relevance, and professional quality embedded throughout the data preparation and experimental pipeline. All quality checks were performed algorithmically to eliminate human error in response validation. The parsing and response validation procedures described in the Methods section (Component Library Neutralization and Validation, Profile Assembly validation, and experimental result parsing) constitute the complete quality control system. Human oversight was limited to designing validation criteria and reviewing final aggregated results. The framework's comprehensive test suite (147 scripts, 41,000+ lines) underwent extensive validation by the authors to ensure technical correctness and methodological integrity.

Originality. This research represents entirely novel work introducing a new methodological paradigm. All components are original, with proper accreditation for all referenced materials and methodologies.

Bias Mitigation. We systematically addressed presentation bias by analyzing positional ordering effects in the experimental results. The Analysis of Presentation Order Bias (Results section) demonstrates that observed signal detection effects reflect genuine content-based discrimination rather than positional artifacts.

Accountability and Transparency. We accept full accountability for all AI-generated content in this research. The AI outputs are fully documented: eminence scores (LLM A/GPT-5), OCEAN personality scores (LLM B/Claude 4.5 Sonnet), neutralized astrological text (LLM C/Gemini 2.5 Pro), and similarity score matrices (seven evaluation LLMs). Complete documentation of tools, versions, parameters, and procedures is provided throughout the article and in the public repository. LLM assistance in manuscript preparation is disclosed in the Author Contributions section.

Broader Impact. This research addresses the replication crisis in psychological science by providing the scientific community with an open-source, automated, and methodologically rigorous framework for investigating complex narrative systems. As an uncompensated research project, this work focuses specifically on researching LLM technology capabilities—a goal integral to the research question itself.

Acknowledgements

The authors wish to thank Vincent Godbout for generously sharing his pioneering thoughts, drafts, and procedures on automated matching tests, which provided a valuable foundation for this work. The authors are independent researchers and received no specific funding for this study.

Open Data and Code Availability

In accordance with the principles of open science and methodological reproducibility (The Turing Way Community, 2022), all data, analysis scripts, supplementary materials, and documentation necessary to reproduce the findings reported in this article are

permanently and publicly available at https://github.com/peterjmarko/llm-narrative-framework.git.

Repository Contents:

- **README**: Quick start guide and framework overview.
- **Replication Guide** (Supplementary Material): Project overview, description of interactive tools and production codebase, complete step-by-step procedures for all three replication paths, including detailed descriptions of the data preparation pipeline, validation procedures, and experiment workflow.
- **Framework Manual**: Technical specifications, data formats, and API references.
- **Source Code**: Complete Python and PowerShell codebase (147 scripts, 41,000+ lines) with comprehensive test suite and 40 technical diagrams.
- **Data Files**: Static datasets for direct replication (34 files total), including:
    - Neutralized component library (CSV format with component IDs and neutralized text)
    - Final subject database (CSV format with biographical and astrological metadata)
    - Raw experimental results (JSON format with trial-level data and model responses)
    - Compiled study-level analysis results (CSV format with summary statistics)
- **Configuration Files**: Exact parameter settings used in the original study.
- **Data Dictionaries**: Complete documentation of variable names, data types, valid ranges, and missing data codes for all datasets.

Example data structures and loading scripts are included to facilitate immediate data access and reuse.

Licensing: The framework is released under dual licensing: source code under GNU GPL v3.0, and data/documentation under CC BY-SA 4.0.

References

Astro-Databank. (n.d.). [Online database]. Astrodienst AG. Retrieved from https://www.astro.com/astro-databank/Main_Page

Astrodatabank Research Tool. (n.d.). [Online tool]. Astrodienst AG. Retrieved from https://www.astro.com/adb-search/

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

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.

Carlson, S. (1985). A double-blind test of astrology. *Nature*, *318*(6045), 419-425.

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), 281–302.

Currey, R. (2022). Meta-analysis of recent advances in natal astrology using a universal effect-size. *Correlation*, *34*(2), 43-55.

Dean, G., & Kelly, I. W. (2003). Is astrology relevant to consciousness and psi? *Journal of Consciousness Studies*, *10*(6-7), 175-198.

Ertel, S. (2009). Appraisal of Shawn Carlson's renowned astrology tests. *Journal of Scientific Exploration*, *23*(2), 125-137.

Eysenck, H. J., & Nias, D. K. (1982). *Astrology: Science or superstition?* St. Martin's Press.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(24), e2305016120.

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

Google. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. Google AI. https://arxiv.org/abs/2403.05530

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

Lewis, J. R. (1994). Southern Hemisphere. In *The Astrology Encyclopedia* (p. 484). Visible Ink Press.

Marko, P. J. (2018). Boomers and the lunar defect. *The Astrological Journal, 60*(1), 35-39.

McRitchie, K. (2022). How to think about the astrology research program: An essay considering emergent effects. *Journal of Scientific Exploration*, *36*(4), 706-716. https://doi.org/10.31275/20222641

McRitchie, K., & Marko, P. J. (n.d.). Is astrology relevant to what consciousness is like? (manuscript in preparation)

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

OpenRouter.ai. (n.d.). [Online API service]. Retrieved from https://openrouter.ai/

Ryder, N. B. (1965). The Cohort as a Concept in the Study of Social Change. *American Sociological Review*, *30*(6), 843-861.

Solar Fire. (n.d.). [Software]. Astrolabe Inc. Retrieved from https://alabe.com/solarfireV9.html

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

van Dongen, N., & van Grootel, L. (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

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://arxiv.org/abs/2206.07682