Detecting a Non-Random Signal in Astrological Descriptions Using a Large Language Model as an Impartial Arbiter

[Author 1], [Author 2]

[Date]

Abstract

Background: Empirical validation of complex, holistic systems like astrology has proven challenging due to methodological limitations. **Objective:** This study introduces a novel, fully automated methodology to test the hypothesis that astrological descriptions contain a non-random signal that correlates with an individual's biography, using a Large Language Model (LLM) as an unbiased pattern-recognition tool. **Methods:** Personality descriptions for a curated database of famous individuals were generated using a deterministic astrological program and programmatically neutralized by an LLM to remove all esoteric terminology. A second, independent LLM was then tasked with matching these descriptions to biographical profiles across varying group sizes (k=4 to 30) in both "correct" and "random" mapping conditions. Performance was measured using lift metrics, which normalize for chance. **Results:** A Two-Way ANOVA revealed that the LLM's performance was statistically significantly higher in the "correct" mapping condition than in the "random" condition. While the effect size was small, this indicates the detection of a subtle but consistent signal. As expected, performance lift also significantly decreased as group size increased. **Conclusion:** The study provides quantitative, reproducible evidence that neutralized astrological descriptions contain a faint but statistically significant signal. This methodology represents a promising new framework for the empirical investigation of complex narrative systems.

Keywords: Astrology, Large Language Models, Methodology, Pattern Recognition, Reproducibility, Open Science

1. Introduction

For centuries, astrology has postulated a meaningful correlation between celestial configurations at the time of birth and human personality. However, empirical validation of this claim has proven notoriously difficult. While landmark studies have attempted to test these claims (e.g., Carlson, 1985), their methodologies and conclusions have been subject to extensive debate (Eysenck & Nias, 1982; Ertel, 2009). Recent reviews suggest that the most promising results have emerged from holistic "whole-chart" correlational tests rather than from attempts to isolate single predictive factors (Currey, 2022; McRitchie, 2022). This modern approach, pioneered in computational form by Godbout (2020), avoids many historical design flaws but highlights the need for even more rigorous, unbiased, and reproducible testing paradigms.

The advent of Large Language Models (LLMs) presents a transformative opportunity to advance this specific line of inquiry. LLMs are powerful, general-purpose pattern-recognition engines that have demonstrated sophisticated capabilities in complex reasoning, pattern recognition, and even personality assessment from text (Google, 2024; Wei et al., 2022). Their ability to process and compare the complex narratives of both biographies and personality descriptions with near-human nuance makes them an ideal instrument for this task. Crucially, they can be deployed as impartial arbiters, executing a

well-defined matching task without prior knowledge of the experimental hypothesis or the theoretical basis of the stimuli.

As a conceptual replication and methodological advancement of prior computational research (Godbout, 2020), this study leverages this novel capability to test a foundational astrological hypothesis: **Do personality descriptions derived from a deterministic astrological algorithm contain a non-random, discernible signal that correlates with the biographical data of the individuals they purport to describe?**

To test this, we employ a novel experimental approach leveraging two distinct LLM applications. First, an LLM is used to programmatically "neutralize" a library of astrological description components. Second, an independent LLM acts as an impartial arbiter to perform a complex "who's who" matching task using profiles assembled from these components. By comparing the LLM's success rate on correctly mapped profiles versus randomly mapped profiles, we can isolate and quantify the presence of any underlying systematic signal. Our primary hypothesis is that the LLM's performance in the "correct" mapping condition will be significantly higher than in the "random" mapping condition, as measured by performance lift over chance.

While the successful detection of such a signal would raise profound questions about consciousness, meaning, and the nature of pattern recognition, the scope of the present study is strictly empirical. We aim to establish a robust, quantitative answer to the foundational question of whether a discernible signal exists. The broader philosophical implications of this finding are taken up in a companion article (Authors, manuscript in preparation).

2. Methods

2.1. Sample Population

The study utilized a curated database of test subjects drawn from the Astro-Databank (ADB) public database (Astro-Databank, n.d.). An initial query using the Astrodatabank Research Tool (Astrodatabank Research Tool, n.d.) yielded 10,378 candidates based on the following primary filters:

- **Biographical Availability**: Subjects were selected from the "Famous: Top 5% of Profession" category to increase the likelihood of comprehensive biographical information being available to the LLM.
- **Data Reliability**: Only subjects with a Rodden Rating of 'AA' (from birth certificate) or 'A' (from memory or a reliable source) were included. Subjects with birth times given only in hours were excluded to avoid rounded data.
- **Ethical Considerations**: The sample was filtered for "Personal: Death" to avoid the need for consent to use personal information.

This list was then subjected to a final programmatic filter to verify a sufficient knowledge base for the LLM. This required that each subject (1) had a corresponding Wikipedia page, and (2) that this page contained an identifiable death date, which served to correct for

occasional inaccuracies in the ADB's "Personal: Death" filter. This final step reduced the sample to 6,193 pre-selected candidates.

To ensure the final sample consisted of well-known and clearly distinguishable figures, the 6,193 candidates were subjected to a final ranking procedure. Four distinct LLMs (o3 mini high, Claude 3.7 Sonnet, ChatGPT 4o, and Gemini 2.0 Flash) evaluated each candidate for both public eminence and OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) traits. The candidates were then ranked using the average eminence score as the primary criterion. This analysis revealed that meaningful distinctions in eminence scores diminished beyond the top 5,000 individuals. To optimize the sample's suitability for the matching task, the final study database was therefore composed of these top 5,000 persons.

2.2. Stimulus Generation and Neutralization

The textual stimuli were generated through a multi-step process designed to ensure that final descriptions were constructed from a consistent, pre-neutralized set of interpretive components.

1. **Component Library Neutralization:** First, the entire library of interpretive delineations within **Solar Fire v9.0.3** (Astrolabe Inc., n.d.) was systematically neutralized. This "cookbook" of text components, which contains descriptions for every possible astrological combination (e.g., planet in a sign, elemental dominances), was processed by **OpenAI's o3 mini**, a model from the GPT-3 lineage (Brown et al., 2020). Each individual snippet was rewritten using the prompt: *"Revise the attached text with the exception of lines marked with an asterisk, which need to remain intact: remove references to astrology and astronomy: shift from second-person perspective to an impersonal, objective, neutral style without referring to specific people; and correct for grammar and spelling. Preserve original text as much as possible while making these revisions."* The lines marked with an asterisk, which were preserved verbatim, served as the unique lookup keys for each component in the neutralized library. This one-time process created a master database of neutralized personality description components.

2. **Placement Extraction:** Second, for each of the 5,000 individuals in the study database, a foundational set of astrological placements was exported from Solar Fire. This structured data included the factors necessary to generate the "Balances" (Planetary Dominance) and "Chart Points" reports, covering elements, modes, and the sign placements of the 12 key chart points (Sun through Pluto, Ascendant, and Midheaven). This specific set of factors was chosen deliberately to test for a primary, non-interactive signal while minimizing the confounding variables that could arise from more complex astrological techniques, such as planetary aspects or midpoints.

3. **Profile Assembly:** Finally, each individual's complete personality profile was programmatically assembled. Their specific set of astrological placements,

exported in the previous step, was used as a key to look up and concatenate the corresponding pre-neutralized description components from the master database. This assembly process resulted in a unique, composite personality profile for each individual, expressed in neutral language, which formed the basis of the stimuli used in the matching task.

2.3. Experimental Design and Procedure

The study employed a 2 x 6 factorial design. The independent variables were:

- ***`mapping_strategy`***: A between-groups factor with two levels: *`correct`* (descriptions were correctly paired with biographical profiles) and *`random`* (descriptions were randomly shuffled and paired).
- ***k* (Group Size)**: A within-groups factor representing the number of subjects to be matched in a given trial, with six levels: 4, 7, 10, 15, 20, and 30.

The core matching task was executed by **Google's Gemini 1.5 Flash** LLM (Google, 2024). For each trial, the LLM was provided with a randomly shuffled list of *k* neutralized personality descriptions and a list of corresponding but independently shuffled *k* names. It was then tasked with independently sourcing the biographical information for each individual before performing the matching. It was prompted with a highly structured request to produce a similarity score matrix. A portion of the prompt is excerpted here: *"You are expected to source the biographies of and any other relevant information about the {k} named people... Please provide your answer only in the format of a table... Each cell... should contain a numerical score from 0.00 to 1.00... Do not include any other text, explanations, or introductions..."*

The experiment consisted of 100 trials per replication, with 30 full replications conducted for each of the 12 conditions (*`2 mapping_strategy levels x 6 k levels`*), totaling 360 complete experimental runs. With 30 replications per condition, this design provided sufficient statistical power (>.80) to detect small-to-medium effect sizes.

The selection of Google's *`Gemini 1.5 Flash`* as the evaluation LLM was the result of a systematic piloting process. A range of models available via the OpenRouter API (OpenRouter.ai, n.d.) were tested for their performance on the matching task, response time, cost-effectiveness, and reliability in adhering to the structured output format. While several high-performing models were considered, *`Gemini 1.5 Flash`* provided the optimal balance of these criteria for the large-scale querying required by this study. To monitor the integrity of the matching process, the LLM was also periodically queried to provide a detailed explanation of its methodology. These introspective checks were reviewed to ensure the model was operating within the intended parameters of the task and not applying external, domain-specific knowledge.

2.4. Dependent Variables and Statistical Analysis

The primary dependent variables were "lift" metrics, which 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.
- **Effect Size (r) and Stouffer's Z-score**: Combined metrics of statistical effect size.

A Two-Way Analysis of Variance (ANOVA) was conducted for each metric to assess the main effects of *mapping_strategy* and *k*, as well as their interaction. Effect sizes were calculated using eta-squared (η^2) to determine the proportion of variance attributable to each factor (Cohen, 1988). Post-hoc comparisons for significant main effects were performed using Tukey's HSD test. The significance level was set at $\alpha = .05$.

2.5. Data and Code Availability

All data, analysis scripts, and supplementary materials necessary to reproduce the findings of this study are openly available in a GitHub repository at [Insert GitHub Repository URL Here].

3. Results

The analysis revealed statistically significant main effects for both *mapping_strategy* and *k* on the most critical performance metrics. The interaction effect (*mapping_strategy * k*) was found to be not statistically significant for the primary lift metrics (e.g., for MRR Lift, $F(5, 348) = 1.13$, $p = .345$). However, a significant interaction was observed for the raw performance metric *Mean Rank of Correct ID* ($F(5, 348) = 2.72$, $p = .020$), indicating that the magnitude of the difference in raw rank between mapping strategies varies with group size. Given our focus on chance-corrected lift metrics, the main effects are of primary interest.

3.1. Main Effect of mapping_strategy

A statistically significant main effect of *mapping_strategy* was found for multiple lift and effect size metrics, consistently showing that the LLM performed better in the *correct* condition than in the *random* condition.

- **MRR Lift**: $F(1, 353) = 6.26$, $p = .013$, $\eta^2 = .015$.
- **Top-1 Accuracy Lift**: $F(1, 353) = 5.17$, $p = .024$, $\eta^2 = .009$.
- **Top-3 Accuracy Lift**: $F(1, 353) = 4.44$, $p = .036$, $\eta^2 = .007$.
- **Effect Size (r)**: $F(1, 353) = 4.95$, $p = .027$, $\eta^2 = .012$.

Figure 1 illustrates the difference in performance lift between the two mapping strategies, showing a small but consistent advantage for the *correct* condition.

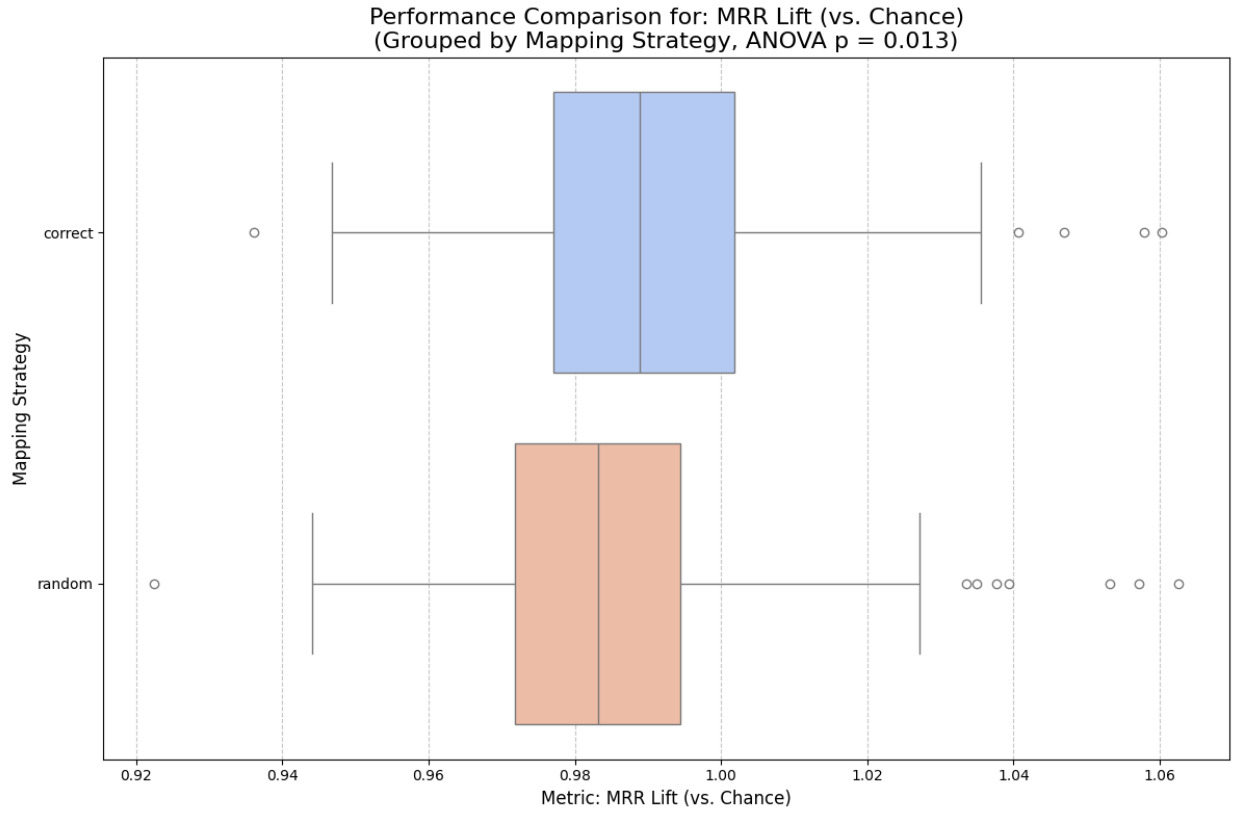

Figure 1: Comparison of MRR Lift (vs. Chance) between Correct and Random mapping strategies.

3.2. Main Effect of Group Size (k)

As hypothesized, *k* had a strong, statistically significant main effect on all lift metrics ($p < .001$ for all). Post-hoc tests confirmed that performance lift systematically decreased as *k* increased. This confirms that the astrological signal, while detectable, is more easily leveraged in less complex choice environments.

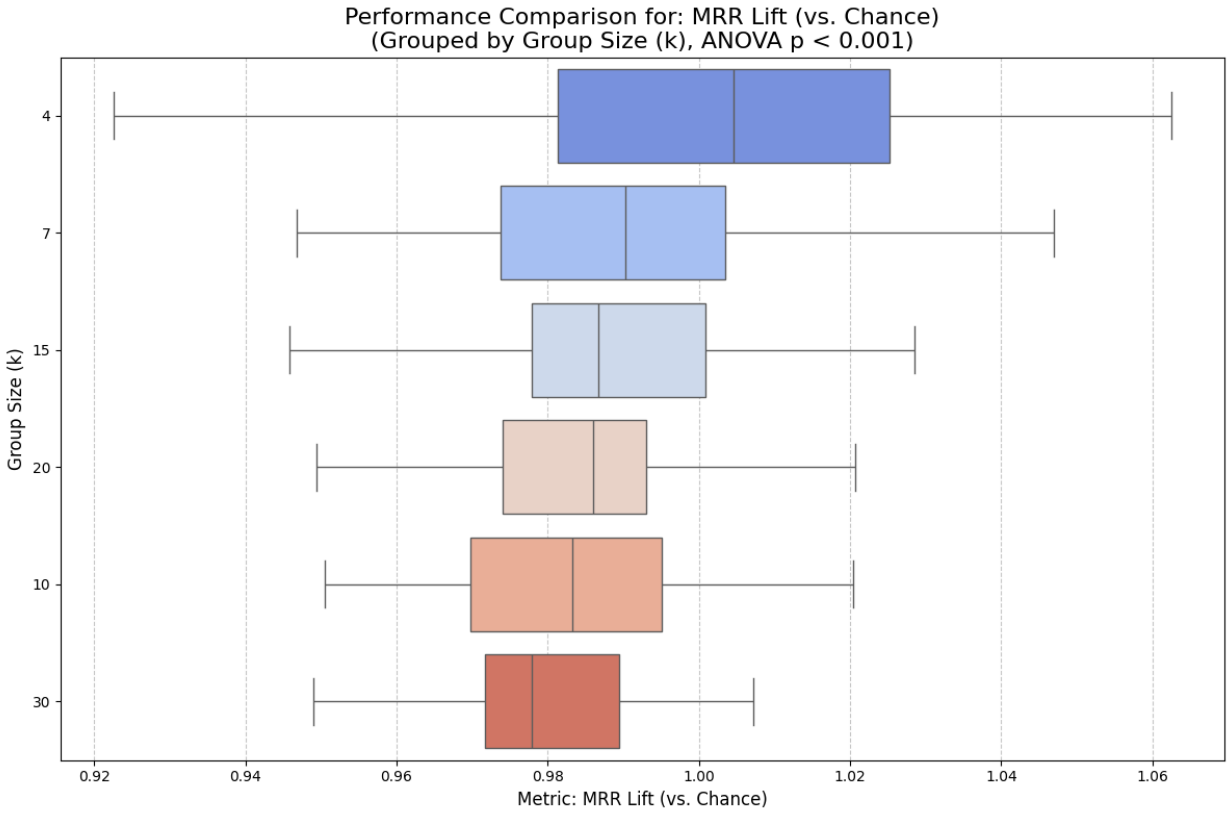

Figure 2: Comparison of MRR Lift (vs. Chance) across different group sizes (k).

3.3. Analysis of Positional Bias

The study also analyzed potential positional biases in the LLM's responses. The ANOVA for *`Top-1 Prediction Bias (Std Dev)`*—a measure of how consistently the LLM preferred certain ranked positions—showed a significant effect for group size *k* ($p < .001$) but not for *`mapping_strategy`* ($p = .357$). This indicates that while group size (*k*) influenced the consistency of the LLM's choices, this behavior did not differ between the correct and random conditions. Importantly, the analyses for *`Bias Slope`* and *`Bias P-value`* showed no statistically significant effects for either *`mapping_strategy`* or *k*, suggesting the absence of a simple linear positional bias in the rankings.

4. Discussion

As a successful conceptual replication and methodological extension of Godbout (2020), the results of this study provide quantitative evidence supporting the hypothesis that

neutralized descriptions generated by an astrological algorithm contain a non-random, discernible signal. The LLM, acting as an impartial pattern-recognition tool, was able to distinguish between correctly mapped and randomly mapped profiles at a statistically significant level. This finding is notable as it overcomes many of the historical methodological hurdles in empirical astrological research. Furthermore, the analysis of positional bias showed no evidence of a systematic linear bias in the LLM's rankings. While the *consistency* of its top choices was affected by group size (k), the model did not appear to favor certain positions in its ranked output, reinforcing the integrity of the core performance metrics.

The subtlety of the detected signal, indicated by the small effect sizes, is an important aspect of this finding. It suggests that the predictive information, while present, is weak and likely insufficient for high-accuracy predictions in practical, high-noise environments. This may be due to several factors: the inherent nature of the phenomenon, the loss of information during the necessary neutralization process, or the constraints placed on the astrological factors used for generation. The significant effect of k reinforces this interpretation; the faint signal is more easily obscured as the number of potential distractors (noise) increases.

Furthermore, the lack of a significant interaction effect between $mapping\_strategy$ and k for the primary lift metrics is an important finding in itself. It suggests that the magnitude of the astrological signal's effect, while subtle, remains relatively consistent regardless of the task's complexity. The signal does not appear to become disproportionately stronger or weaker as more distractors are introduced.

Beyond the specific findings related to astrology, the primary contribution of this work may be the demonstration of a novel methodological framework. This LLM-driven, fully automated pipeline offers a robust and impartial template for the empirical investigation of other complex, narrative-based systems that have historically resisted quantitative analysis. Potential applications could include testing the coherence of Jungian archetypal analyses, evaluating the consistency of diagnostic systems from traditional Chinese medicine, or validating thematic coding in qualitative social science. Moreover, in an era where the behavioral sciences are grappling with a "reproducibility crisis" (Open Science Collaboration, 2015), the framework presented here offers a path forward. By automating the entire experimental lifecycle and removing the human experimenter from the core data generation and analysis loop, this approach eliminates a significant source of potential bias and allows for direct, verifiable replication by other researchers.

It is important to consider alternative explanations for these findings. One common critique of personality descriptions is that they are susceptible to the "Barnum effect," where vague, generalized statements appear to apply to almost anyone. While this effect is potent in single-case validations, the comparative, forced-choice design of this study inherently controls for it. The LLM was not asked to simply validate a single profile, but to determine the *best* fit from a set of up to 30 distinct options, a task that requires significant differentiation. Another potential confound is the well-documented "birth season" effect,

where the time of year of one's birth correlates with certain life outcomes. However, the signal tested in this study is far more complex than a simple seasonal variable; it is derived from a multi-factorial system of 12 distinct chart points and their distribution across various divisions. It is therefore unlikely that a broad seasonal effect could account for the specific, nuanced signal detected by the LLM, though this remains a potential area for future control.

Ultimately, this study was designed to answer a single, empirical question: is there a detectable signal? The results indicate that the answer is yes. The profound question of *what it means* for a non-conscious, algorithmic system to detect a faint but significant pattern within a symbolic framework traditionally associated with human meaning-making is a philosophical one. This deeper inquiry, which explores the implications for our understanding of consciousness and pattern recognition, is the subject of a companion analysis (Authors, manuscript in preparation).

4.1. Limitations and Future Directions

A primary limitation of any research utilizing closed-source LLMs is the "black box" nature of the models. Despite the neutralization of input texts, it is theoretically possible that the LLM could recognize the underlying patterns as astrological in origin and use its own latent knowledge of astrology to "cheat" the test, rather than detecting the signal from first principles. While creating a custom-trained LLM with no astrological knowledge was beyond the scope of this study, this potential confound was addressed via introspective checks. The LLM's self-reported explanations consistently described a method of reverse-engineering objective traits from the biographies to find the best match, with no indication that it was knowingly applying astrological principles. Furthermore, when explicitly queried, the LLM stated that it does not consider astrology a reliable source of biographical information, suggesting it was not actively trying to apply such a framework. While not definitive proof, this provides evidence that the model was operating as the impartial pattern-recognition tool intended by the experimental design.

This study has several other limitations. First, it relies on specific LLMs for neutralization and evaluation; results may differ with other models. Second, the sample was restricted to famous individuals, whose widely known biographies may introduce confounding variables. Future research should replicate this methodology with different astrological techniques (e.g., aspects, midpoints, different house systems, inclusion of more complex factors), different LLMs, and non-public-figure populations to assess the generalizability of these findings. Exploring the impact of the neutralization process itself would also be a valuable avenue of investigation. Finally, the astrology expert system itself is specific to this study, providing a further avenue to explore.

5. Conclusion

This study successfully deployed a novel, automated, and objective methodology for testing a core hypothesis of astrology. The findings indicate the presence of a faint but statistically significant signal within neutralized astrological descriptions, detectable by a

sophisticated, impartial AI arbiter. This work does not validate astrology as a whole, but it challenges the null hypothesis that its outputs are purely arbitrary. It provides a robust, reproducible framework for future empirical investigations and establishes a firm factual basis for the subsequent philosophical inquiry into consciousness and symbolic systems explored in its companion article.

Author Contributions

- **[Author 1 Name]:** Conceptualization; Methodology; Software; Formal Analysis; Investigation; Writing – Original Draft; Writing – Review & Editing.
- **[Author 2 Name]:** Conceptualization; Supervision; Writing – Review & Editing.
- *(Please edit roles as needed based on CRediT taxonomy)*

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.

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.

Appendix: Settings for "Balances Report"

The calculation of Solar Fire's "balances report" (planetary dominances) utilized the default weighting system, with one key modification. Based on exploratory trials, the weights for the generational planets (Uranus, Neptune, and Pluto) were set to zero to isolate more individualized factors. The specific "weight-points" assigned were as follows:

- **3 points:** Sun, Moon, Ascendant (Asc), Midheaven (MC)
- **2 points:** Mercury, Venus, Mars
- **1 point:** Jupiter, Saturn
- **0 points:** Uranus, Neptune, Pluto

Dominance within each astrological category (e.g., elements, modes) is automatically determined by the program through a multi-step calculation:

1. A "total score" (TS) is calculated for each division (e.g., the element 'fire', the mode 'cardinal') by summing the "weight-points" of all chart points located within it.
2. An "average score" (AS) is then determined for the category by averaging the TS values across all its constituent divisions.
3. Two thresholds are established using this AS and predefined ratios: a "weak threshold" (WT) calculated with a "weak ratio" (WR), and a "strong threshold" (ST) calculated with a "strong ratio" (SR):

- $WT = AS * WR$
- $ST = AS * SR$

4. Finally, a division is classified as 'weak' if its TS was below the WT, or 'strong' if its TS was greater than or equal to the ST.

The interpretive output of this process is the resulting list of 'strong' and 'weak' classifications for each division, which is then used for profile assembly.

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., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (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.

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

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

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.

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

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

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/

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

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://arxiv.org/abs/2206.07682