Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 16, No. 2, 2026

Interpretations of Tarot Card Spreads by AI: Predictive Versus Introspective Questions

Zixi Cynthia Wang *

Shenzhen College of International Education, Shenzhen, Guangdong, China
*** Corresponding author Email:** Cynthia.Wang@stu.scie.com.cn

Abstract: To explore whether AI-generated interpretations of Tarot card spreads are different when predicting external events versus introspection on one's own thoughts. Thirty items made up the balanced test questions for prediction and verification; after querying, card-specific analysis results would be given as a reference. For each of the question-spread pairs, this study obtained five answers from DeepSeek and five from GPT-4, along with a human-divination score classified as "no", "partly yes" or "yes". Convert all responses into numbers (0, 1 or 2) and then calculate the average chatbot's response per question; next, perform a two-factor ANOVA on questions divided by groups and people to answer levels. ANOVA showed a significant main effect of the human answer level ($p=0.0017$), indicating that AI scores are correlated with human certainty at this point. However, a one-year follow-up of the "partly-yes" group found that there was still a significant difference ($P < 0.01$). Predictive questions had an average score of 1.58 compared to 0.82 for introspection; the mean difference reached 2.13, indicating it was highly pronounced (Cohen's D). Based on these results, although AI interpretations generally conform to human judgement, question type affects their output under an ambiguous human reference. Predictive Questions may trigger stronger affirmations due to bias in the training data. A new method for measuring sensibility in divination Systems and its impact on the behaviour of AI Advisory System through questions framing is demonstrated here.

Keywords: Tarot Interpretation; Large Language Models; Divination; Question Framing; ANOVA; Human-AI Alignment.

1. Introduction

For a long time, scholars have been interested in examining the accuracy or inaccuracy of tarot prediction from psychological perspectives; furthermore, semiotics were used to study how meaningful things arise from symbols and circumstances within tarot card readings[1]-[3]. However, these customs have rarely received any examination through computational means. As continuous development of the application scenarios for large-scale language models to offer advisory and diagnostic functions, more research directions related to this topic have attracted scholars' attention.

This study poses a relatively straightforward query. Does an AI-produced interpretation of Tarot spreads vary, depending on whether the issue is focused outside or inside (that is, will one get a job; am I on track)? In other words, do variations in the correspondence between AI results and humans exist for different questions? And when will such discrepancies occur?

The majority of the data are categorical or text-based; there were thirty such questions in total, along with corresponding three-card spreads, humans' answers, and ten AI responses. After preprocessing, all categories of responses were converted to numerical form; specifically, 0 represented "no", 1 indicated a response of "partially yes", and 2 denoted "yes"; the scores were then averaged over questions to obtain the initial continuous dependent variable. The project output includes statistical Comparisons, including the outcomes of ANOVA and T-tests for quantifying the relationship between AI interpretation and question design as well as human judgments.

Using psychology's understanding of divination, semantic analysis of symbolic systems, and computational methods from statistics, this study builds an evaluation framework for assessing the behaviour of artificial intelligence in situations with constructed meaning. When does the AI model make a judgment error? That is, under different circumstances of the user's personal condition within specific scenarios, how should users be encouraged or educated? It may not meet such demands.

2. Background

There are generally two types of research directions for understanding divination systems such as Tarot; psychological explanations, and symbolic or semiotic analyses. Based on both traditions, this paper introduces the computing power of this paper.

Psychological approaches are most often illustrated through research into the Barnum Effect and Cold Reading. Dutton offered a basic explanation for how tarot readers, astrologers, and others use calculated guesses combined with generalised characterisations to make the process appealing to the public [1]. Building on the Barnum Effect, which makes people tend to believe vague statements that are universal enough but still apply to them individually, Ivtzan reviewed the relevant literature on tarot card divination and compared it to psychological factors causing misjudgment or subjectively verifying outcomes; therefore, it cannot be proven whether there is an effect of supernatural causes [2]. As a matter of fact, in order to establish the argument based on real evidence, one should be similarly adaptable. But at the same time, it considers tarot to be a trick rather than an orderly system of symbols beneath it.

A different method views divination as a cognitive-semiotic system. Based on educational bibliography, Semetsky's research suggests that using tarots can help people reflect inwardly upon themselves and their values [3]. Recently archived works on Dust interpret the Yijing as a kind of cognitive system, derive an invariant law for modelling

changes in The Xici and connect these phenomena with more general questions regarding how change manifests itself philosophically [4]. This tradition reflects that people recognise the underlying cause of the system and are inclined to make sense in this way. It has the problem of excessive emphasis on qualitative research and avoiding quantitative verification.

Thirdly, a line of research examines divination in terms of indigenous knowledge systems and computation techniques. Olagunju et al. represent the basic ifá divination symbols algebraically, and present that the primary Odudibia are contained in a definite binary form [5]. Omoregie argues that ifa is a kind of indigenous decision support systems with code outputs that need to be interpreted by using narratives, contexts and community knowledge [6]. This method has its own differences from others; that is, the algorithm itself does not have a context for generating meanings outside this dialogue.

Moreover, in recent years, several computational model studies on the semantic organisation of symbolic systems have emerged. Garg et al. found that, according to the changes and associations of words' cultural stereotypes over time [7]. Kozlowski et al. used embeddings to analyze cultural meaning structures in class discourse [8]. These studies inform the present approach by demonstrating that computational tools can uncover latent patterns in symbolic language.

This study links these traditions. Based on psychological theory, this study takes an objective view and conducts empirical research instead of mysticism. According to the semiotics tradition, this study takes an interest in how divination works as a structured symbolic system. From a computational perspective of symbolic knowledge, this study obtains the methodological rule that such systems can be modelled, quantified and validated. What distinguishes this study from other research is that it uses statistics, including conducting one-way ANOVA and T-test analyses on people's interpretive sensitivity towards Tarot cards; by providing these typical philosophical or psychological problems with an objective evaluation method based on data.

3. Dataset

A new dataset was developed for this study to explore whether there are differences in AI's interpretation results of Tarot card spreads under different questioning scenarios, specifically two types, predictive questions (e.g., about future outcomes; asking if something will go well) and introspective questions (asking about internal situations or feelings; like self-assessment). Mainly consisting of categorical and textual data, it has been converted to numerical form after being subjected to transformation.

3.1. Data Type and Size

In each row of the dataset, there is the question text, the three-card spread, the question group (predictive or introspective), a human-divination answer, ten chatbot responses (including five deep learning models: one from DeepSeek and four from GPT-4; average score calculation results) and their standard deviation values. Table 1 shows the functions adopted in this study.

That is to say, the whole dataset contains 30 sets × 14 items = 420 pieces in all. Given that it was an exploratory investigation aimed at identifying patterns in the data, there was no need to partition the sample for use as both training and test sets. A total of 30 samples were chosen to identify whether they could be used as reference objects in the study on interaction patterns between three elements: question-type, human-AI judgment, and AI-output.

3.2. Data Collection and Preprocessing

The questions were curated from real-world tarot inquiry sources, including online forums and social media platforms, to preserve ecological validity. Each question was paired with a three-card spread drawn from the Rider-Waite tarot deck, selected to represent a variety of suits and major arcana. For every question-spread pair, five independent responses were obtained from DeepSeek and five from GPT-4 using a consistent prompt that instructed the model to interpret the cards in the context of the question and output one of three categorical answers: "yes", "partially yes", or "no". The human reference answer was recorded on the same scale.

Preprocessing involved several steps. First, all text entries were converted to lowercase to ensure consistency. The categorical answers (both human and chatbot) were then mapped to numerical values: "no" -> 0, "partially yes" -> 1, and "yes" -> 2. A new feature, Average Chatbot Score, was calculated as the row-wise mean of the ten chatbot response columns, providing a continuous measure of the AI's central tendency for each question. Minor formatting inconsistencies, such as extra spaces or inconsistent punctuation, were corrected manually during cleaning.

3.3. Feature Descriptions and Meanings

Table 1. Feature descriptions and meanings

Feature	Type	Description	Meaning
Question	Text	The specific question asked	Defines the context for interpretation
Cards	Text	The three tarot cards drawn	Provides the symbolic input for the reading
Group	Categorical	Predictive vs. introspective	Tests whether question type affects AI output
Answer	Categorical (0, 1, 2)	Human diviner's standard	Serves as the reference point for comparison
Chatbot responses (DeepSeek / GPT-4)	Categorical (0, 1, 2)	Individual AI responses	Capture variability across models and runs
Average Chatbot Score	Continuous / float	Mean of all 10 chatbot responses	Primary dependent variable for statistical analysis

Features that could help answer the central research question of this paper: How do AI Interpretations Differ Across Question Types and Their Relationship with Human Judgments? Table 1 shows the variables and their contents.

Verify whether there are differences in the prediction results compared with self-expression in different situations of Group functionality. Answer function can be used for the comparison between AI output and humans at different confidence levels. Multiple chatbots running can be used as an indicator of their operational reliability; meanwhile, average chatbot score is calculated from one series to eliminate fluctuations for pattern identification.

Therefore, in total, there are 30 balanced training examples; text and category variables have been converted to numerical forms for this study. Together they constitute the core direction of the research problems as follows. Question type, human judgment; response pattern by AI.

4. Methodology

To test whether there are differences in the accuracy of AI-generated interpretation results for predictions versus self-reflection based on tarot card spreads, as well as whether they align with people's subjective judgment, this study applied data preprocessing, statistical tests, and charts in Python. As it is an exploratory study that does not require prediction, no train/test split was conducted, but all 30 data points were analysed together.

4.1. Data Preprocessing

The raw data was loaded from a Google Sheet into a pandas DataFrame. Upon initial inspection, there were some mixed-case categories in certain categorical fields, such as "Yes" and "PARTIALLY YES". Uniformity was ensured by converting all text contents in specific cells to lowercase using the str.lower() function. Then, by mapping the categories "no", "partially yes", and "yes" to the numbers 0, 1, and 2 respectively, each column containing a human response and ten random chatbot replies was converted to these three categories. This mapping enabled data quantification through pandas.map using a dictionary. Any rows containing missing or invalid data were carefully checked for elimination; no rows were deleted.

A new variable called "Average Chatbot Score" was created after encoding to calculate the row-wise average of all ten chatbot response columns. As the core dependent variable, this continuous variable captured the central tendency of the AI interpretation among each spread-paired question answers and smoothed over runs-of-error variance. Finally, the resulting DataFrame had a total of 30x14=420 cells, which were categorically marked in preparation for further examination.

4.2. Error Variable Construction

An error variable was computed to quantify the extent to which the AI's answer differed from that of humans in this case. The error is defined as follows.

$$Error = (Answer - Average\ Chatbot\ Score)\text{^}2$$

This continuously takes into account the size of disagreement between the AI and the human reference, without considering its direction. Finally, this study compared the error differences in predictive and reflective groups through an independent-samples t-test. Given that there is no difference in the quality of AI interpretation between questions 1 and 2, the mean errors are expected to be close.

4.3. Statistics Collection

To determine whether Question Type and Human Judgment affect AI Scores, a two-way ANOVA was conducted. The group was considered as an indicator of category type at two points (predominantly predictive or intuitive); answered similarly categorically divided into three grades (0 = no, 1 = partially yes, 2 = yes).

The average chatbot score served as the dependent variable. Both the main effects and the group-by-answer interaction in this model were included. This selection made it possible to determine whether there were differences in the AI score among different question types, whether these discrepancies matched people's assessments; and see if this difference was stable for humans' various answer levels.

The ols method from the statsmodels.formula.api package was used to perform the ANOVA test. The formula of this model is as follows.

Average_Chatbot_Score ~ C(Group) + C(Answer) + C(Group):C(Answer)

Fit the model, and then generated its ANOVA table using anova_lm from statsmodels.stats.anova. Performed residual diagnosis via normal Q−Q plot and homoscedasticity check, reasonably confirming compliance with the assumptions of ANOVA; there was no serious violation found.

To test the following hypotheses proposed by the ANOVA, an independent-t-test was conducted using scipy.stats. ttest_ ind. Two of these tests compared average chatbot score for group differences at each point with respect to "partially yes"; the second compared errors in both groups respectively.

4.4. Implementation

The analysis of all parts was carried out using a Google Cloud Platform-based collaboration and repeat-reliability-oriented Jupyter Notebook environment.

Pandas was mainly used to manage and clean the data; NumPy was employed for mathematical computations; Matplotlib and Seaborn were utilised for visualisation purposes; Statsmodels carried out ANOVA; and Scipy.stats conducted a t-test. All code can also be found including comments and other relevant information within the attached notebooks.

5. Results and Discussion

A two-way ANOVA was employed to determine whether there were differences in predicting or introspecting among different groups at both levels of humans' responses, namely none, partly correct, and fully right. An independent-samples t-test was then performed to examine group discrepancies exclusively among those who answered partially in agreement ("partially yes").

These approaches were selected to answer the research questions directly: ANOVA is used for a general test of effects; the focus on t-test or other methods can reveal subtle patterns not evident in an all-at-once examination.

5.1. Two-Way ANOVA Results

A two-way ANOVA was performed with Group (predictive, introspective) and Answer (0 = no, 1 = partially yes, 2 = yes) as independent variables and Average Chatbot Score as the dependent variable. The ANOVA results are summarized in Table 2.

Table 2. Two-way ANOVA results for average chatbot score

Source	Sum of Squares	df	Mean Square	F	p-value
Group	0.164	1	0.164	2.92	0.097
Answer	1.205	2	0.602	10.72	0.0017
Group x Answer	0.215	2	0.107	1.91	0.172
Residual	1.350	24	0.056		

The main effect of Answer was statistically significant (p=0.0017); that is to say, there were differences in chatbot scores among various levels of human judgments on average. In terms of descriptions, the score "yes" was higher than that of "no"; partially also appeared in middle positions. This indicates that the AI's response tends to have a similar polarised effect as the human reference; when the human's answer is clear-cut, so are those of the AI.

There was no significant main effect of Group (p = 0.097), indicating that at the individual level of each answer range, there was no overall difference in chatbot scores between predictive and introspective questions. Similarly, there was no significant difference between Group and Answer; thus, the influence on questions of different answers was also the same. Fig. 1 shows that the lines of the predictive and introspective groups are still approximately parallel, indicating that there is little interaction; the increment from No to Yes is fairly similar among all groups.

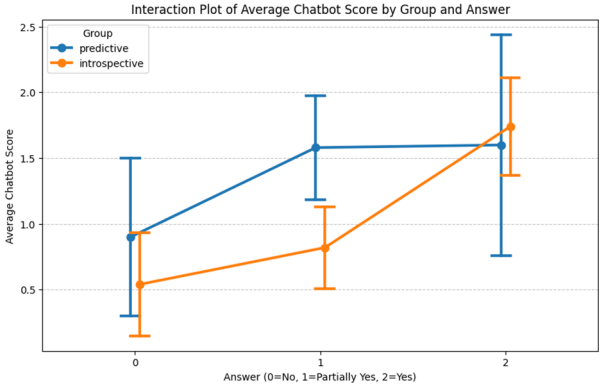

Fig 1. Interaction plot of average chatbot score by question group and human answer level. Error bars indicate variation within each condition

As a result, since the dataset is balanced between various people's answers, this supports the interpretation that AI output is correlated with human assessment results rather than chance factors in this study.

5.2. Analysis of "Partially Yes" Answers

While the overall ANOVA revealed no significant Group effect, it was hypothesised that differences might arise only under conditions of ambiguous human judgment ("Partially Yes") might present separately. This was tested by removing the ten questions that were answered "partially yes" by humans to ensure an equal distribution for predictive and reflective types, with five items each. An independent-samples t-test compared the average chatbot scores of these two groups.

t-test showed that there was a statistically significant difference (t(8) = 3.372, P=0.010). Predictive items had a mean score of 1.58, with a standard deviation of 0.37; introspective items scored only 0.82, with a standard deviation of 0.33. This shows that when the human reference is ambiguous, in prediction situations, the AI has a greater tendency to respond positively than it does for self-examination; the effect size (Cohen's d) was at a considerable level, thus there is practical significance.

As shown in Figure 2, the predictive group obtained a much higher mean score compared with the introspective group; this indicates statistical significance for "Partially Yes" answers.

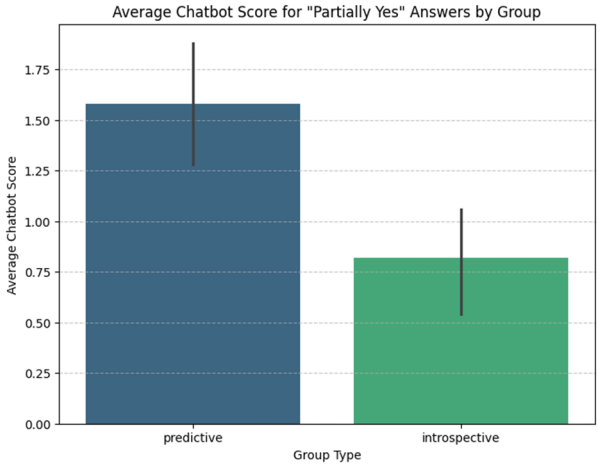

Fig 2. Bar plot comparing average chatbot scores for predictive and introspective questions with a "partially yes" human answer. Error bars indicate uncertainty around the group means.

5.3. Error Analysis

To quantify the magnitude of AI-human disagreement, the squared error was computed (Answer - Average Chatbot Score)^2 for each question. An independent-samples t-test comparing this error between predictive and introspective groups yielded a non-significant result (t(28) = 0.84, p = 0.41). The mean error for predictive questions was 0.21 (SD = 0.19) and for introspective questions 0.18 (SD = 0.15). Thus, although the direction of AI scores differed in the partially yes subset, the overall error magnitude was similar across groups. This suggests that the AI's performance, measured as deviation from human judgment, is not systematically better or worse for one question type; rather, the bias manifests as a systematic shift toward more affirmative responses for predictive questions when human judgment is ambiguous.

5.4. Discussion

Based on the above conclusions, there are some related meanings regarding artificial intelligence understanding Tarot Reading. Firstly, there is a clear significant first-order effect of humans' answers on the output scores; the system responds accordingly to this polarity of human-answer levels. The results show that the AIs have obtained some association with

card combinations and question answers. The average score has increased from "no" to "yes"; there is some extent of semantic match among them.

Secondly, due to its absence of a global group effect, in general terms, the artificial intelligence has treated both predictive and reflective queries equally. It may be that the prompt did not clearly distinguish among them, or the card meanings do not have direct ties to their respective questions' types. Nevertheless, there is a considerable variation among those who responded "yes"; some people's responses were vague; therefore, the AI predicted that they would answer "yes". A possible reason might be that predictive questions (e.g., questions about future results) are more likely to receive positive responses in the model's training data; however, introspective questions may be linked to less overtly affirmative or negative language expressions. Another reason could be that the types of card spreads selected by both groups were slightly different in terms of meaning; however, this was uncontrolled.

There is an obvious difference between groups according to the t-test; it cannot be considered insignificant. Question presentation may affect people's understanding of artificial intelligence, with uncertainty being a factor among them. This indicates the limitations of applying AI for advice-giving scenarios: the users' requests for information about future development have higher agreement scores than those seeking self-reflection; although it aligns well with human judgments, its extent may still vary somewhat.

Based on this, the reason for differences among people's evaluations has been determined to be that people generally agree with AI on most things. It can be seen that although the AI has achieved high accuracy in general cases, when making predictions under uncertain conditions, it tends to overestimate more frequently compared with other types of errors.

5.5. Limitations and Future Directions

However, there are some deficiencies in the above-mentioned data. There are limitations in terms of generalisation with the small number of participants; there were only 30 items in total for all subjects, and those contributing significantly to the most notable findings consisted of merely ten questions. Because of this small number, there might be some sensitivity to single extreme cases. Although there were differences among the three groups, with more data from further experiments, the results will be better supported later on.

Subjectivity in the responses from people can also introduce sources of bias. The human reference answers were given by a few tarot experts, and different readers may label the same question-card combination in various ways; thus, there is no definitive answer but only interpretation. In the future, multiple human responses may be collected from different items to enhance the reliability of the reference criteria.

Variability in AI responses is another drawback. Responses were averaged over ten times to remove some fluctuations; however, each time an individual response deviated from others—the same model would give varied results when faced with the same input. It might be due to the random nature of stochastic large language models; there was no measurement or control over the temperature setting that affected output variability.

Card selection was non-random; there were particular selections made. Card spreads were selected from various suit cards and the high-level Aces of Major Arcana; however, it cannot be guaranteed that each combination meets this standard. Some of these cards may naturally favour Yes/No responses and introduce bias into the group comparison. In fact, randomisation methods should be used to assign cards for various questions at random to avoid results caused solely by differences among the categories of questions.

The responses generated by prompts under various circumstances are somewhat unchanging; nevertheless, there may be slight variations due to differences in wording such as "Will it happen?" and "Think about it", which might affect later model output results accordingly. Although the AI provided indifferent responses and did not give any positive results, it was still insufficient to examine whether there are differences among individuals in their reactions after modifying text word meanings.

In addition, this was confirmed by statistics. The ANOVA assumption of normal distribution in terms of means is consistent after visual inspection (acceptable). However, in smaller groups compared with larger ones, it may be less dependable. The t-test also assumes that variances in different groups are equal; however, after checking, they appeared similar at first glance.

There is an absence of specific control over card meanings, which could affect how the AI responds differently than what this study has tracked. Cards that have a high frequency of occurrence among some question groups may cause biased results. In a future research project, researchers should add card-level semantic control or balanced handling procedures to separate the impact of card content from question setting.

In the next research, the following should be improved: More data will provide a more robust ability to make judgments; different directions of collecting standard cards provide ample material for research. Multiple humans should be asked to re-verify the classification; if there is consistency in all cases, and it remains clear-cut when answering simple "yes" / "no", then this phenomenon may extend beyond the framework used for label interpretation. Bridging the cards' spread among groups through a randomised design will diminish confounding from card selection, and investigating various prompt styles (neutrality vs directive) may shed light on how question frames affect model behaviour. Mixed-effects models could offer a richer analytical tool for use here.

Although this research does not have perfect theoretical support to explore how artificial intelligence creates symbols under various framing conditions. The discovery that the prediction question elicited a higher proportion of positive answers under ambiguous human judgment suggests other questions on how the artificial intelligence model internally constructs query framing to affect its behaviour during application.

6. Conclusion

Whether there are differences in the interpretation results of AI for predictive versus introspective tarot card questionnaires? This study created a balanced set of 30 questions with the same three-card distribution; ten artificial intelligence responses were collected per question (five DeepSeek and five GPT-4) and sought out human references from tarot masters. Following the pre-processing of the data, after calculating the results of a two-way ANOVA with a focus on the "partially yes" particular group, in addition to analysing another newly created error variable.

Key findings are as follows. There is an obvious main effect of human answer level ($p=0.0017$) indicating that AI's score is significantly related to the correctness of humans' answers; when humans said "Yes", AI's score was higher than when they said "No". The learning of the AI model has achieved certain correspondence between card combinations and answer directions. Secondly, although there was no significant global difference among the various question types, when human judgment was uncertain, it could be found that predictive questions obtained relatively high averages in prediction scores ($m=1.58$) compared to introspective ones ($m=0.82$); moreover, this difference reached a considerable extent—i.e., Cohen's $d=2.13$. The error analysis showed that there was not much difference in terms of the extent of AI-human disagreement among different people; thus, the direction of bias is opposite to it.

There is no global group effect, and there are substantial differences between groups; thus, it can be concluded from this that Question-Framing Effects have their greatest impact when people find the correct answer unclear. Predictive questions may elicit more positive responses due to the frequent occurrence of ideal answers in training data related to a forward-looking orientation. Practically speaking, those who want to know about external events might get more positively worded system-provided explanations compared with the others.

Several limitations may exist; particularly the short trial duration per person, a relatively small number of jurors, and reliance on manual cards. In the future, a substantial amount of data is needed with multiple annotators' assessments to train the models. Although there are some limitations, this study presents a computational method for investigating the divination system through psychological theory of meaning-creation, semantic analysis on symbolic structure, and contemporary statistical approaches. Through quantitative analysis of how artificial intelligence interprets tarot under various question directions, this paper is the first to explore how these models create meaning within a context that lacks an objective right answer.

References

[1] D. L. Dutton, "The cold reading technique," Experientia, vol. 44, no. 4, pp. 326-332, 1988. DOI: 10.1007/BF01961271.

[2] I. Ivtzan, "Tarot Cards: A Literature Review and Evaluation of Psychic Versus Psychological Explanations," Journal of Parapsychology, vol. 71, no. 1, pp. 139-149, 2007. URL: https://www.parapsychologypress.org/jparticle/jp-71-1-139-149.

[3] I. Semetsky, The Edusemiotics of Images: Essays on the Art~Science of Tarot. Rotterdam, The Netherlands: SensePublishers, 2013. DOI: 10.1007/978-94-6209-055-2.

[4] M. Dust, "The Yijing as Cognitive System: The Principle of Invariance in Phenomena Dynamics Through Dual Perspectives," PhilArchive, manuscript, 2026. URL: https:// philarchive. org/rec/DUSTYA.

[5] A. S. Olagunju, A. A. James, E. O. Adeyefa, and F. L. Joseph, "Algebraic characterization of Ifa main divination codes," Scientific African, vol. 20, Art. no. e01729, 2023. DOI: 10. 1016/ j.sciaf.2023.e01729.

[6] U. Omoregie, "What Ifa, an Indigenous binary knowledge system can teach us about AI," LSE Impact Blog, Aug. 22, 2025. URL: https://blogs.lse.ac.uk/ impactofsocialsciences/ 2025/08/22/what-ifa-an-indigenous-binary-knowledge-system-can-teach-us-about-ai/.

[7] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," Proceedings of the National Academy of Sciences, vol. 115, no. 16, pp. E3635-E3644, 2018. DOI: 10.1073/ pnas. 1720347115.

[8] A. C. Kozlowski, M. Taddy, and J. A. Evans, "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings," American Sociological Review, vol. 84, no. 5, pp. 905-949, 2019. DOI: 10.1177/0003122419877135.