

Optimizing AI-based assessment in history education: The impact of prompt engineering on scoring performance in a morphologically rich language

Yunus Özdemir^{a*} , Emine Aşçı^a 

^a Bolu Abant İzzet Baysal University, Bolu, Türkiye

Suggested citation: Özdemir, Y. & Aşçı, E. (2026). Optimizing AI-based assessment in history education: The impact of prompt engineering on scoring performance in a morphologically rich language. *Journal of Educational Technology and Online Learning*, 9(3), 374-396.

Highlights

- Prompt engineering is more critical than model architecture for achieving high rater reliability in AI-based assessment.
- Structured prompts with explicit rubrics significantly mitigate the "generosity bias" frequently observed in Large Language Models.
- The study demonstrates the efficacy of LLMs in scoring open-ended history responses within the specific context of Turkish, a morphologically rich language.
- Advanced models like Gemini 2.5 and Claude 3.7 achieve human-level scoring consistency when guided by detailed evaluation criteria.

Abstract

This study investigates the effectiveness and prompt sensitivity of leading large language model (LLM)-based AI tools (ChatGPT, Gemini, Claude, Deepseek) in grading open-ended exam questions in an undergraduate history course (Atatürk's Principles and History of Revolution), compared to human evaluators. A comprehensive dataset comprising 72 distinct open-ended responses (collected from 24 students) was scored by both the course instructor and seven AI models using five different prompts with varying levels of detail. These prompts ranged from a basic zero-shot instruction to progressively more structured designs: expected topic headings, a weighted criterion-based rubric, partial coverage with language proficiency, and relative (norm-referenced) scoring. Correlation and statistical analyses (paired samples t-test, Wilcoxon signed-rank test) revealed that while models showed low agreement with the human evaluator when using unstructured, basic prompts (zero-shot), they achieved high agreement ($r > .80$) when provided with structured prompts and explicit rubrics, particularly in the cases of Gemini and Claude. The findings highlight that for effective AI-based assessment, prompt design and the definition of criteria are more critical than model selection in mitigating "generosity bias." Generosity bias here denotes the models' systematic tendency to assign higher scores than the human rater; the qualitative evidence indicates that it stems primarily from the models rewarding fluent, lengthy, and well-structured answers even when their factual content is incomplete, a tendency that explicit rubrics substantially reduced. Designed as an exploratory case study, these results provide significant empirical evidence regarding the optimization of AI as an assistive assessment tool, specifically within the context of Turkish, a morphologically rich language.

Article Info: Research Article

Keywords: *AI in Education, Natural Language Processing, Morphologically Rich Languages, History Education*

1. Introduction

Open-ended questions play a crucial role in assessing student achievement in education. Students are asked to construct an answer rather than just labeling the given answer as simply correct or not: in this way, higher-order thinking skills (critical thinking, creativity, and so forth) are assessed (Brown, 2004). Indeed, the extant literature emphasizes multiple skills and competencies assessed by open-ended questions that cannot be adequately measured by multiple choice test types (Badger & Thomas, 1992; Cooney et al., 2004). Within a higher education context, traditional examination types largely serve to assess students' grasp of complex concepts and the ability to work with that knowledge. In history education, it's been

* Corresponding author. Division of Instructional Technology, Department of Educational Sciences, Faculty of Education, Bolu Abant İzzet Baysal University, Bolu, Türkiye.

e-mail address: yunusozdemir@ibu.edu.tr

Doi: <http://doi.org/10.31681/jetol.1882892>

Received 5 Feb 2026; Accepted 19 Aug 2026

ISSN: 2618-6586. This is an open Access article under the CC BY license.



observed that open-ended questions stimulate students to articulate their own detailed explanations and interpretations of historic events and to work towards deep understanding. This instructional approach in courses enriched with content, such as Atatürk's Principles and History of Revolution, can strengthen students' critical thinking skills and promote deep engagement with the content matter. It should also be emphasized that Atatürk's Principles and History of the Turkish Revolution is a compulsory common course taken by all undergraduate students in Turkish universities; consequently, improvements in how this course is assessed have an unusually wide institutional reach. The assessment model examined here is likewise transferable to other verbally oriented subjects, such as literature and the social sciences, in which open-ended responses constitute the primary evidence of learning. International policy documents similarly underline the centrality of open-ended tasks and higher-order thinking skills in contemporary curricula and call for their responsible support through digital technologies (OECD, 2023; UNESCO, 2023).

Although open-ended questions can be very educative, they also represent some challenges to the assessment process (Jauhiainen & Guerra, 2024). Scoring such responses is also a time-consuming process and may be influenced by the attitude of the scorer, which could pose issues of rater reliability. At the same time, variations in the creativity of writing of individual students, or even the appearance of their work can affect scoring which sometimes result in subjectivity and inconsistency (Bektaş & Kudubeş, 2014). While open-ended items are pedagogically superior, they suffer from inherent psychometric vulnerabilities, specifically "rater drift" due to fatigue and the 'halo effect.' Unlike traditional Automated Essay Scoring (AES) systems that require extensive training data, LLMs offer a potential solution as a scalable, "zero-shot" assessment tool that can function as a consistent second rater, thereby enhancing inter-rater reliability without the prohibitive costs of model training. This differs from multiple-choice tests, which are popular because they are low-cost, easy-to-grade, and believed to be relatively objective. These tests have been criticized, however, as they may be limited in what they can assess with regard to deeper, as opposed to surface, thinking, when students are usually forced to select a single correct answer (Brown, 2004). This is a serious issue in the area of educational assessment. While open-ended questions are needed to allow a comprehensive evaluation of learning outcomes, they raise major issues in terms of a fair and reliable assessment. Nowadays, new technological methodologies that can help and support the evaluation are becoming popular.

The use of artificial intelligence (AI) applications for education has grown explosively and has just begun to become ripe with various possibilities in the last decade. AI systems have shown success in areas such as adaptive learning systems, intelligent tutoring systems, and student retention prediction (Zawacki-Richter et al., 2019). The introduction of large language models (LLMs) has driven a revolution in educational approaches, where AI-aided systems have brought novel opportunities from a pedagogic as well as a learner point of view (Kasneci et al., 2023). Advanced language models like ChatGPT have been used for numerous educational purposes, including student question-answering and creative writing. Thus, these models open new avenues and present challenges (Morjaria et al., 2024). The use of tools in examination tests is a topic of intensive research, due to its capability of shortening the time-period for testing and the capability of producing objective results. For example, traditional automatic written assessment systems are already implemented and they have been proved to achieve similar levels of consistency as human examiners do in standardized tests (Page, 1966; Dikli, 2006). In contrast, the emergence of first generative AI models over the past few years has led to the anticipation that prevailing models would enable richer considerations of open-ended responses, owing to the fact that they can read and write text at a human level (Holmes et al., 2023).

The growing use of AI in assessment also raises important ethical considerations that frame this study. Recent scholarship emphasizes that AI-assisted scoring must satisfy the principles of transparency (the ability to explain how a score was produced), fairness (freedom from bias against particular groups, languages, or answer styles), accountability (clearly assigned human responsibility for grading decisions), and data privacy (protection of student work and identity) (Holmes et al., 2023; UNESCO, 2023; Yan et al., 2024). A systematic review by Yan et al. (2024) found that most LLM-based educational innovations remain insufficiently transparent and rarely include human-in-the-loop safeguards, underscoring the need

for empirical validation studies such as the present one. Within this framing, AI-assisted assessment is conceived here not as a replacement for the instructor but as a supervised assistive tool: it can substantially reduce instructors' grading workload for open-ended items while final grading authority, oversight, and responsibility remain with the human educator. The design of this study, in which the instructor's criterion-referenced evaluation serves as the reference standard and all student data were anonymized, operationalizes this ethical, human-centred approach.

In fact, early experiments show that large language models can produce results that resemble those from human evaluators in the task of exam marking. Flodén (2025) also used the ChatGPT 3.5 model to mark three high education final exams. AI scores were compared to instructor scores in a data set consisting of 463 responses. The findings showed that 70% of the grades ChatGPT gave were within $\pm 10\%$ of the instructor's grades, whilst 31% were within $\pm 5\%$. Nonetheless, anecdotal evidence suggested that ChatGPT also erred, tending to be somewhat more generous (often referred to as leniency bias) in its value assignments, less critical towards itself, than had been the human raters. As in the previous example, a study including short answers related to medical training concluded that ChatGPT and an experienced educator performed comparably (ranges from $r = 0.60$ to $r = 0.70$) (Morjaria et al., 2024). Based on the study, AI could evaluate responses as effectively as a single expert evaluator, given the right guidelines. They show that AI has an important role in dealing with open-ended responses. The results showed that the evaluators were amazed by how similar the AI's scores were to their own scores (Flodén, 2025).

However, the usefulness and consistency of using AI for assessments are still being discussed today. Because the decision-making processes of AI models are not fully transparent, it is possible that their results will show biases from the data they have trained with (Flodén, 2025). Consequently, tools that use AI must undergo thorough validation testing before being used in open-ended exams. It is also worth noting that unlike STEM fields where answers are often binary (correct/incorrect), History as a discipline involves high interpretive subjectivity. Therefore, investigating whether prompt engineering can align AI with human grading in such a nuanced domain serves as a critical stress-test for the models' semantic processing capabilities, holding significant promise for the broader humanities. This study aims to confirm whether current artificial intelligence (AI) tools are suitable for grading university-level open-ended test questions and to explain their potential use in evaluating examinations. In this situation, the instructor and AI tools both looked at how students responded to open-ended questions about Modern Turkish History. The relationships between AI and human scores were studied by doing correlation analysis and the statistical significance of any score differences was determined.

This research is designed as an exploratory case study focusing on prompt sensitivity. While the student sample size is targeted, the methodological depth is achieved through a multi-dimensional analysis involving 840 distinct data points (24 students $\times$ 5 prompts $\times$ 7 models) at the composite-score level. Throughout this article, three data levels are distinguished to avoid ambiguity: 72 written responses (24 students $\times$ 3 questions), 2,520 item-level AI scoring events (24 $\times$ 3 $\times$ 5 $\times$ 7), and 840 composite-level data points obtained when each student's three item scores are averaged into a single composite score (24 $\times$ 5 $\times$ 7). All inferential statistics are conducted on the composite scores, with the student as the unit of analysis ($N = 24$), allowing for a robust evaluation of model-prompt sensitivity in a specialized academic domain. This study also examines the psychometric robustness of LLM-based scoring outside of Anglocentric contexts. By testing model alignment in a morphologically rich and agglutinative language like Turkish, we aim to investigate whether LLM-generated scores maintain construct validity and scoring stability under complex linguistic constraints, serving as a stress-test for the tool's cross-lingual generalizability.

In line with this purpose, the study is guided by the following research questions:

- **RQ1.** How do different prompt structures, ranging from zero-shot instructions to explicitly criterion-based rubrics, affect the scoring alignment between AI models and the human evaluator?
- **RQ2.** Do AI models with different architectures (ChatGPT, Gemini, Claude, and DeepSeek) produce consistent results when evaluating historical texts written in Turkish, a morphologically rich language?

- **RQ3.** To what extent do AI models exhibit a systematic scoring bias relative to the human evaluator, and how is this bias affected by the level of detail specified in the prompt?

2. Literature

2.1. Open-Ended Assessment Approaches

Open-ended questions allow students to demonstrate their knowledge by formulating their own responses. It is based on constructivist learning theory. Due to this approach, students are expected to take part in thinking about questions, to present their reasoning and to state their views openly (Linn & Miller, 2005). Being able to answer open-ended questions lets students show their own backgrounds, understanding and opinions in their answers. Because of this, students are supported in applying what they know by constructing responses rather than merely repeating what they have learned. Open-ended questions are considered an important method to gauge students' creative thinking, problem-solving, language skills and their involvement in certain classes (Kubiszyn & Borich, 2024). Free-response assessments are meaningful for assessing student learning, according to constructivist theory, due to helping students display advanced thinking skills.

Cognitive theories are also used by specialists to explain the effect of open-ended judgments on learning. Bloom's taxonomy shows that teachers should use open-ended questions to check if students can think analytically, combine ideas or evaluate what they have learned. If used within a constructivist approach such open-ended assessment also looks at students' ability to transfer and use what they have learned, as well as explain it in their own words. Here, assessment is treated as an important part of learning, since students' replies are looked at as outcomes and also feedback which helps continued learning.

2.2. Principles of Validity in Measurement and Evaluation

Ensuring accurate and valuable outcomes from measurements depends on the validity-based design in evaluation. Messick (1995) explains that validity covers more than test questions and includes a broad set of information and reasons backing up how results can be used. Within this system, evidence from content, construct and criterion validity supports validity of the process. The match between what is measured by the assessment and its intended learning goals helps guarantee the overall validity of the assessment. The way students respond is analyzed for correctness as well as their thinking, how they organize their answers and their writing style which all play a part in the area we want to measure (AERA, APA, & NCME, 2014).

Previous validity issues in scoring student responses are now being joined by those due to the use of artificial intelligence tools. If artificial intelligence gives inconsistent scores to similar inputs or does not clearly explain its scoring system, this can threaten the construct validity of the system (Williamson et al., 2006). Also, since the model automatically finds patterns in data, this can make it challenging to evaluate answers that include different cultural or linguistic elements. In this situation, validity involves the correctness of results as well as making sure the assessment is fair to all. But if these tests are not designed with care, they might end up favoring some students more than others, creating new forms of educational inequality (Popham, 2020). As a consequence, checking how similar AI-generated and human evaluator scores are is regarded as a quality standard for this study.

2.3. Prompt Sensitivity of Large Language Models (LLMs)

Large language models (LLMs) are designed using large quantities of data so they can produce responses that make sense in various situations. Such models are remarkably sensitive to the prompt given by the user which is an important impression. A number of studies show that issuing the same instructions in various ways often creates widely different responses from AI-powered LLMs (He et al., 2024). This indicates that context can shape what models say, since their answers are not the same each time. An important difference becomes clear when we say, "evaluate" versus "evaluate as an expert historian, scoring the meaning and way the text is presented." It points to the need for prompt engineering when doing evaluations (Brucks and Toubia, 2025).

Due to the way these prompts vary, people must be cautious about the reliability and validity of the LLM assessments. According to Pinto et al. (2023), ChatGPT produced different scores for the same student responses depending on the question format and the results could also be interpreted differently. Both identifying the model as either an expert, a teacher or an exam grader and explaining the evaluation standards had a major effect on the outcomes. Consequently, it is proposed that employing multiple prompts in evaluation may engender more reliable outcomes than artificial intelligence analysis predicated on a solitary prompt, with respect to evaluation consistency (Carlson and Burbano, 2025). In this study, five distinct prompts were administered to each AI tool, and prompt sensitivity was systematically analyzed.

3. Methodology

3.1. Research Design and Participants

This research is designed as an exploratory classroom implementation study utilizing a correlational framework. While the statistical analysis maintains the student as the unit of analysis ($N=24$) to ensure methodological independence, the study technically processes a comprehensive dataset of 2,520 distinct AI scoring events ($24 \text{ students} \times 3 \text{ questions} \times 5 \text{ prompts} \times 7 \text{ models}$). Consistent with the terminology defined in the Introduction, these 2,520 item-level scoring events correspond to 72 written responses and reduce to 840 composite-level data points once each student's three item scores are averaged; all hypothesis tests reported below are based on these composite scores ($N = 24$). This high-volume data generation allows for a robust technical evaluation of prompt sensitivity and model performance despite the focused sample size.

The participants consisted of 24 undergraduate students enrolled in the “Atatürk’s Principles and History of Turkish Revolution” course at a state university. Of the participants, 16 were female and 8 were male, and their ages ranged from 19 to 22. In terms of year of study, 22 of the students were in their second year, one was in the first year, and one was in the fourth year of undergraduate education. To ensure ecological validity, the course instructor’s evaluation was established as the ‘ground truth’ for this study. Given the instructor’s deep contextual knowledge of the specific curriculum and student progress, this approach prioritizes pedagogical consistency over the inter-rater agreement typically sought in standardized testing environments. Consequently, the study focuses on the AI’s alignment with this pedagogical standard rather than a consensus between multiple human raters.

3.2. Data Collection Process

Data collection was done during the 2024-2025 spring semester as part of the course’s midterm exam on Atatürk’s Principles and History of the Revolution. The instructor carefully designed three big questions to guide the examination. The main topics in the investigations were major topics in Turkish Revolutionary History, and the point was to check students’ knowledge and ability to understand significant events and concepts. During the exam, each student provided detailed written answers to the three questions. After finishing the examination, the instructor looked through the students’ answers. Each question in the evaluation got a score between 0 and 100 according to the course learning aims. The maximum possible total score was 300 points. Each question was deliberately scored on a separate 0-100 scale, rather than as a fraction of a single 100-point exam, for three reasons. First, each question targeted a distinct learning outcome of the course, and an independent 0-100 criterion-referenced scale allows the quality of each answer to be judged in its own right. Second, a 0-100 range provides sufficient granularity for the weighted, percentage-based rubric used in Prompt 3, which would be distorted if compressed into a 30-35 point sub-scale. Third, this design yields comparable score distributions across the three questions, which is a prerequisite for the composite-score analyses reported in the Findings. What was expected in each answer (the required content elements, correct use of historical terminology, chronological accuracy, and coherence of expression) is defined in the question-specific criteria presented in the Appendices, and the overall data collection and scoring workflow is visualized in Figure 1. These scores were taken as a reference in the research.

Following the evaluation of the students’ answers by the instructor, the same written answers were also evaluated by artificial intelligence-based tools. To this end, the students’ answers were transferred to a

digital environment, and each answer text was submitted to the selected AI tools. The study employed a range of prominent artificial intelligence tools that were based on large language models from that period. The following software programs were utilized: ChatGPT (GPT-4o, GPT-4.1, and o3 versions), Google Gemini 2.5 Pro, Claude 3.7 Sonnet, and DeepSeek (versions V3 and R1). The selection of AI models was predicated on architectural diversity and reasoning capabilities. Specifically, Claude 3.7 Sonnet was included for its advanced coding and reasoning alignment, Gemini 2.5 for its multimodal processing, and DeepSeek R1 for its reinforcement learning-based logic. This variety ensures that the results represent a cross-section of current large language model (LLM) paradigms rather than a single company's technology.

To make the rationale for model selection explicit, Table 1 summarizes the main similarities and differences among the seven models employed in the study.

Table 1.

Comparison of the AI models employed in the study

Model	Developer	Model type / architectural emphasis	Rationale for inclusion
ChatGPT (GPT-4o)	OpenAI	Multimodal general-purpose flagship model	Widely used baseline in AI-assisted grading research
ChatGPT (GPT-4.1)	OpenAI	Instruction-following optimized successor model	Enables testing the effect of incremental version updates within the same vendor
ChatGPT (o3)	OpenAI	Dedicated reasoning (chain-of-thought) model	Represents the reasoning-model paradigm
Claude 3.7 Sonnet	Anthropic	Hybrid reasoning model with strong instruction alignment	Advanced reasoning and rubric-following capability
Gemini 2.5 Pro	Google	Multimodal model with a very large context window	Multimodal processing and long-context capacity
DeepSeek V3	DeepSeek	Open-weight mixture-of-experts general model	Represents the open-weight LLM paradigm
DeepSeek R1	DeepSeek	Open-weight reinforcement-learning-based reasoning model	Reinforcement learning-based logic; cost-accessible alternative

Note. Model characteristics reflect the versions publicly available at the time of data collection (2024-2025 spring semester).

Each artificial intelligence tool was provided with five distinct prompts, meticulously designed to assess the student's response to the inquiry. For each prompt, the AI tool was tasked with evaluating each student's response according to the established criteria and assigning a score within the range of 0 to 100. This process was repeated for all AI tools, and the evaluation scores assigned by the AI for all student responses were collected. To ensure the consistency of the outputs obtained from the AI tools, the prompt statements given to each AI tool were standardized. The evaluation process for all AI tools was conducted under the supervision of researchers and within the same timeframe.

Several procedural controls were applied during the AI scoring sessions. First, the data-sharing and model-improvement options of the AI tools were disabled, so that no student response could be used for model training and no session could influence another. Second, for Prompts 1-4, each student's answer to a single question was submitted individually, in a separate and independent chat session; this prevented carry-over effects, since a model evaluating several answers within the same conversation could be influenced by its previous messages. Third, all evaluations except the relative-scoring condition were carried out back to back within the same time period, because even short intervals between sessions could alter a model's response tendencies and undermine the equality of conditions. Only the relative-scoring condition (Prompt 5) required the full set of 24 responses to be presented within a single context session, as the scoring was defined relative to all answers in the group; preliminary checks conducted before this step confirmed that the combined text remained within the models' context-window (token) limits, and no truncation or omitted response was observed in any scoring session.

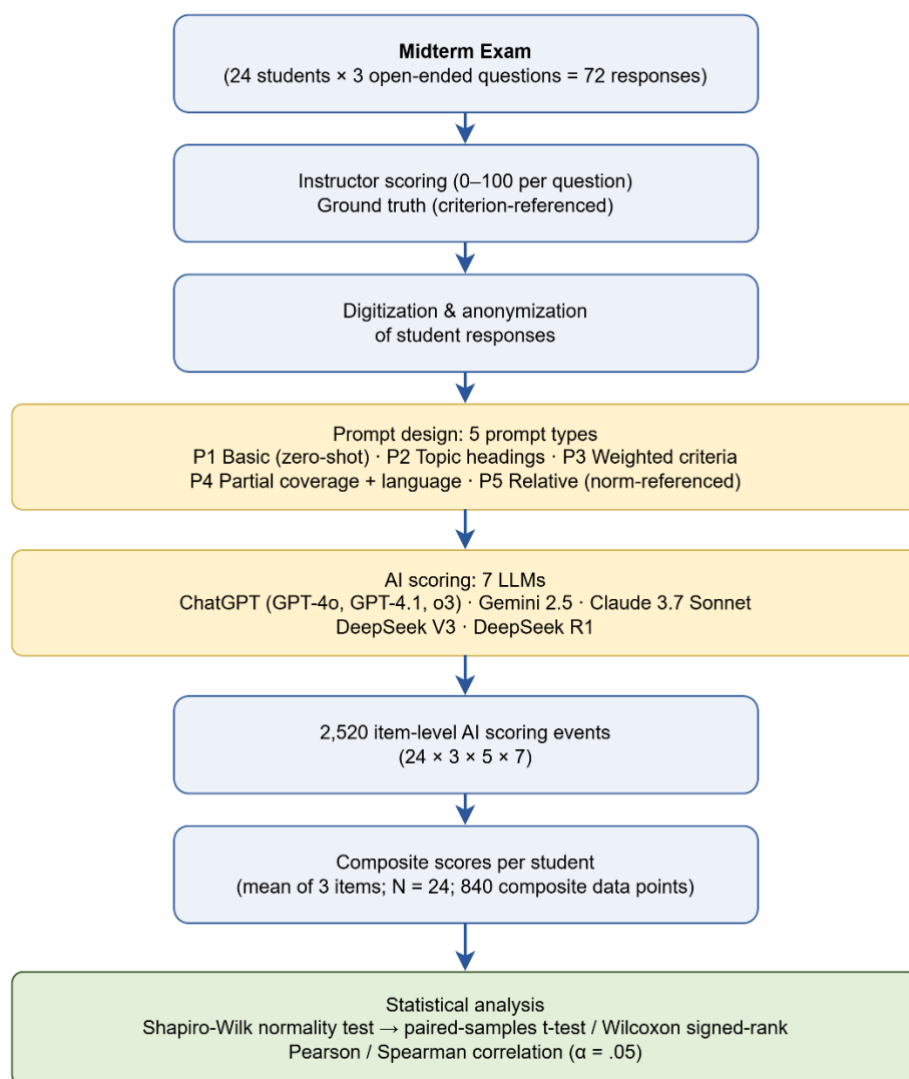


Fig. 1. Flow diagram of the data collection, prompt design, and scoring workflow.

3.3. Data Analysis

The human and artificial intelligence scores obtained in the study were examined using quantitative analysis techniques. The distribution of the scoring data was examined using the Shapiro-Wilk test. When the normality assumption was not met, the Wilcoxon signed-rank test was applied. When it was met, the paired samples t-test was applied. The results of the normality tests are presented in Appendix 7.

The strength of the relationship between the instructor's scores and the scores produced by AI tools was determined through the implementation of correlation analysis. Pearson correlation coefficient was used for normally distributed data, while Spearman's rank correlation was employed for non-normal distributions. For each AI tool and each prompt, the correlation coefficient between the scores given by the relevant tool and the human scores was calculated separately. Although the dataset comprises 72 distinct responses (24 students × 3 items), statistical analyses were conducted at the student level (N=24) using composite scores. A composite score was calculated for each student by averaging the scores obtained from the three items. This methodological approach was adopted to mitigate item-specific variance and measurement error, prioritizing the reliability of the overall student achievement score over single-item performance. The statistical significance of the correlation coefficients was then tested at the $\alpha = 0.05$ level.

All statistical analyses were conducted using Jamovi software (The jamovi Project, 2024). In addition, Python statistical libraries were used by the researchers for the purpose of visualization.

3.4. Ethical Principles and Reliability

Prior to conducting the research, the necessary ethical committee approvals were obtained. The Human Research Ethics Committee of the relevant university reviewed the study protocol and granted approval. To ensure data confidentiality, the student responses were coded and anonymized to exclude any personal information. In the case of student responses, the identification data was removed, and only the academic content was shared. All data obtained in the study were used solely for scientific analysis and were not shared with third parties.

3.5. Prompts and Rationale

In this study, a comprehensive evaluation was conducted to assess the capacity of artificial intelligence models to evaluate open-ended questions. To this end, five distinct prompts were utilized, varying in content and instruction levels, to evaluate the performance of these models. The primary objective of the selection process was to methodically assess the model's capacity for scoring sensitivity and its responsiveness to criteria. Methodologically, the experimental design represents a controlled progression from 'Zero-Shot' instructions (Prompt 1) to 'Constraint-Based' and 'Chain-of-Thought' analogues (Prompts 3, 4, and 5). This stratification allows for a granular analysis of how increased prompt specificity directly mitigates algorithmic stochasticity and scoring hallucinations. The following section presents the selected prompts and their respective justifications:

1. Basic Prompt (Simple Instruction):

The request for scoring was elementary in nature. "Please assign a score between 0 and 100 to these answers." The objective of this study was to ascertain the performance of the models in evaluating answers in the absence of guidance or supplementary information, thereby ascertaining their inherent tendency.

2. Prompt based on the expected topic headings:

The subjects were provided with prompts that delineated the subjects to be addressed in their responses. They were also asked to assign a score to each response based on the provided headings. The objective of this study is to assess the sensitivity of the models to the prompt in the context of specific content expectations and to evaluate their performance in content-based scoring.

3. The prompt based on a set of criteria and a percentage:

The content, information, language, expression, chronology, consistency, and other criteria were determined based on expert opinion and assigned specific percentages, which were then fed into the model. The request for separate scoring for each criterion was granted. This enabled the evaluation of the models capacity to implement multi-dimensional and weighted scoring schemes.

4. Prompt based on partial coverage and language proficiency:

While providing all expected topics, full marks are given if only a certain part of these topics is addressed correctly and consistently, and if the language and expression are sufficient. The objective of this study is to examine the models' pursuit of flexibility and consistency in their responses and their methodology for evaluating responses that meet the minimum proficiency level.

5. Prompt for relative (norm-referenced) scoring:

All responses were to be evaluated in relation to each other. The primary objective of this prompt is to assess the capacity of models to assign scores to responses in a relative, rather than an absolute, manner, relative to other responses within the group, and to discern between differences. During this process, considering the context window limits, all 24 student responses were presented to the model within a single context session. The model was instructed to read all responses and rank them relative to one another through a holistic evaluation approach.

In Appendix 6, the original Turkish text of the prompts used is presented, along with English translations that serve to enhance the clarity of the material. This diversified prompt approach has enabled a systematic

analysis of AI models' sensitivity to different scoring paradigms, including absolute and relative scoring, content focus, multi-criteria evaluation, and minimum proficiency.

4. Findings

In this section, the alignment and consistency between the human evaluator (instructor) and AI models in scoring open-ended history questions are analyzed. The analyses are presented under three sub-headings: statistical differences between score means, inter-rater correlation coefficients, and the effect of prompt types on model performance.

4.1. Analysis of Score Differences Between AI and Human Evaluators

The study examined whether the scores assigned by seven different AI models using five distinct prompt types differed statistically from those of the human evaluator. Depending on the normality distribution of the data, either the Paired Samples t-Test or the Wilcoxon Signed-Rank Test was applied. The comparative difference analysis results for all prompt types and models are presented in Table 2.

Table 2.

Analysis Results of Score Differences Between Human and AI Models Across All Prompt Types

Model	Prompt 1 (Basic) (p)	Prompt 2 (Topics) (p)	Prompt 3 (Criteria) (p)	Prompt 4 (Partial) (p)	Prompt 5 (Relative) (p)
ChatGPT o3	<.001	<.001	<.001	<.001	<.001
ChatGPT 4.1	0.295	<.001	<.001 *	0.003	0.148 *
ChatGPT 4o	<.001 *	0.006	<.001 *	0.001	<.001
Claude 3.7	0.001 *	0.474 *	0.115 *	0.066 *	0.011 *
Deepseek R1	<.001 *	<.001	<.001 *	<.001 *	<.001
Deepseek V3	<.001 *	<.001	<.001 *	<.001 *	<.001 *
Gemini 2.5	<.001 *	0.021 *	<.001 *	0.005 *	0.021

* The Wilcoxon signed-rank test was employed for data sets that did not demonstrate a normal distribution.

Table 2 reveals significant variations in model performance depending on the prompt structure. Regarding the Basic Prompt (Prompt 1) configuration, the majority of models produced scores that were statistically significantly different ($p < .05$) from the human evaluator. The notable exception was ChatGPT 4.1, which showed no statistically significant difference ($p = .295$), suggesting an inherent alignment with human scoring in unstructured tasks.

However, as the prompts became more specific and structured, Claude 3.7 emerged as the most consistently reliable model. In contrast to other models that fluctuated, Claude 3.7 demonstrated no statistically significant difference from the human evaluator across all structured prompt categories: Expected Topic Headings (Prompt 2) ($p = .474$), Criteria-Based Evaluation (Prompt 3) ($p = .115$), and Partial Coverage (Prompt 4) ($p = .066$). This indicates that Claude 3.7 is exceptionally responsive to detailed instructions and rubrics, producing score distributions equivalent to the instructor in these contexts.

Finally, in the Relative Scoring (Prompt 5) context, ChatGPT 4.1 was the sole model to align with the human evaluator ($p = .148$). These findings suggest that while Claude 3.7 excels in criteria-referenced scoring, ChatGPT 4.1 shows superior alignment in norm-referenced (ranking-based) assessments.

4.2. Inter-Rater Reliability and Correlation Findings

Spearman/Pearson correlation analyses were conducted to determine the degree of alignment between the scoring tendencies of the AI models and the human evaluator. The correlation coefficients (r) for the different prompt types are summarized in Table 3.

Table 3.

Correlation Coefficients Between Human Evaluator and AI Models

Model	Prompt 1 (Basic) (r)	Prompt 2 (Topics) (r)	Prompt 3 (Criteria) (r)	Prompt 4 (Partial) (r)	Prompt 5 (Relative) (r)
ChatGPT o3	.56	.65	.76	.63	.70
ChatGPT 4o	.82	.64	.84	.76	.71
Claude 3.7	.73	.83	.82	.79	.70
Deepseek R1	.68	.74	.65	.68	.76
Deepseek V3	.66	.62	.63	.75	.74
Gemini 2.5	.77	.82	.82	.81	.81
ChatGPT 4.1	.75	.76	.78	.78	.68

The correlation findings presented in Table 3 reveal that model stability varies significantly depending on the prompt type. In Basic Prompt (Prompt 1), ChatGPT 4o achieved the highest correlation ($r = .82$), contrasting with other models that showed moderate alignment.

The introduction of structured criteria generally strengthened the reliability of top-tier models. In the Expected Topic Headings (Prompt 2) scenario, Claude and Gemini demonstrated exceptional consistency, achieving correlations of $r = .83$ and $r = .82$, respectively. Similarly, in the Criteria-Based (Prompt 3) evaluation, ChatGPT 4o reached the highest overall correlation observed in the study ($r = .84$), followed closely by Claude and Gemini ($r = .82$).

The Gemini model, on the other hand, exhibited the highest consistency across the board, maintaining strong correlation coefficients ranging between .77 and .82 across all five prompt types. This indicates that while some models fluctuate based on instructions, others maintain a stable ranking capability regardless of the prompt structure.

4.3. Qualitative Observations of Scoring Discrepancies

Beyond the statistical indices, the score sheets were also examined qualitatively to identify where and why the models diverged from the instructor. Two recurrent patterns emerged. First, in the unstructured conditions the models tended to reward the structural volume and formal register of an answer rather than its factual completeness. A concrete illustration was observed in the responses concerning the Lausanne Conference (Question 3): one student produced a lengthy and rhetorically coherent essay that received high scores from the models (e.g., 85/100 from ChatGPT 4o), whereas the instructor assigned 60/100 because the answer omitted the Mosul Question, a key learning outcome of the relevant lecture. Second, the models rarely classified subtle historical inaccuracies as critical errors unless the rubric explicitly instructed them to do so. These observations indicate that the generosity bias quantified in the preceding sections is driven primarily by the models' tendency to conflate linguistic fluency and essay length with factual completeness, a pattern that structured, criterion-referenced prompts visibly suppressed. Since the scoring protocol was deliberately designed to keep the outputs of the seven models standardized and comparable, the models were not asked to provide detailed written explanations for their scores; these observations are therefore based on score-level divergences. A systematic, fine-grained analysis of the models' error types (for instance, whether a historical term was misinterpreted because of its morphological inflection) accordingly falls outside the scope of the collected data and is proposed as a direction for future research in the Limitations section.

4.4. Summary of Key Findings

Based on the data obtained from the difference tests and correlation analyses, a summary of the models demonstrating the most consistent performance for each prompt type is presented in Table 4.

Table 4.

Summary of Key Findings

Prompt	Models Showing the Best Correlation	Models with No Statistical Difference
1	ChatGPT 4o ($r=.816$, $p<.001$) Gemini 2.5 ($r=.775$, $p<.001$) ChatGPT 4.1 ($r=.746$, $p<.001$)	ChatGPT 4.1 ($p=.295$, $p>0.05$) t*
2	Claude 3.7 ($r=.833$, $p<.001$) Gemini 2.5 ($r=.821$, $p<.001$) ChatGPT 4.1 ($r=.758$, $p<.001$)	Claude 3.7 ($p=.474$, $p>0.05$) W*
3	ChatGPT 4o ($r=.842$, $p<.001$) Claude 3.7 ($r=.825$, $p<.001$) Gemini 2.5 ($r=.818$, $p<.001$)	Claude 3.7 ($p=.115$, $p>0.05$) W*
4	Gemini 2.5 ($r=.813$, $p<.001$) Claude 3.7 ($r=.790$, $p<.001$) ChatGPT 4.1 ($r=.779$, $p<.001$)	Claude 3.7 ($p=.066$, $p>0.05$) W*
5	Gemini 2.5 ($r=.818$, $p<.001$) Deepseek R1 ($r=.761$, $p<.001$) Deepseek V3 ($r=.744$, $p<.001$)	ChatGPT 4.1 ($p=.148$, $p>0.05$) W*

* t: Paired samples t-test; W: Wilcoxon signed-rank test. The test applied is determined based on whether the relevant data set shows a normal distribution.

The holistic analysis presented in Table 4 identifies the most valid model-prompt combinations, defining validity as the balance between high correlation (r) and statistical equivalence ($p > .05$) with the human evaluator.

- **Prompt 1 (Basic):** While ChatGPT 4o demonstrated the strongest correlation ($r = .816$), ChatGPT 4.1 emerged as the most "human-like" scorer in terms of absolute grades, being the only model with no statistically significant difference ($p = .295$).
- **Prompt 2 (Topics):** Claude 3.7 achieved the optimal performance, recording the highest correlation ($r = .833$) and demonstrating no statistically significant difference ($p = .474$) from the instructor's scores. Gemini 2.5 also showed a strong correlation ($r = .821$) in this category.
- **Prompt 3 (Criteria) & Prompt 4 (Partial):** Although ChatGPT 4o ($r = .842$) and Gemini 2.5 ($r = .813$) showed the highest correlations for Prompts 3 and 4 respectively, Claude 3.7 was the only model to maintain statistical equivalence to the human evaluator in both scenarios ($p = .115$ and $p = .066$). This highlights Claude 3.7's robustness in handling detailed assessment criteria without succumbing to significant scoring bias.
- **Prompt 5 (Relative):** Gemini 2.5 exhibited the strongest ranking capability ($r = .818$), but ChatGPT 4.1 was the sole model to align with the human evaluator's scoring distribution statistically ($p = .148$).

In conclusion, the summary findings suggest that Claude 3.7 offers the most consistent validity across structured prompts (Prompts 2, 3, and 4), successfully balancing high alignment with accurate grade distribution. Meanwhile, Gemini 2.5 consistently performs well in correlation metrics across most prompt types.

To help readers interpret these overarching patterns at a glance, Figure 2 visualizes the correlation coefficients reported in Table 3 together with the model-prompt combinations that showed no statistically significant difference from the human evaluator.

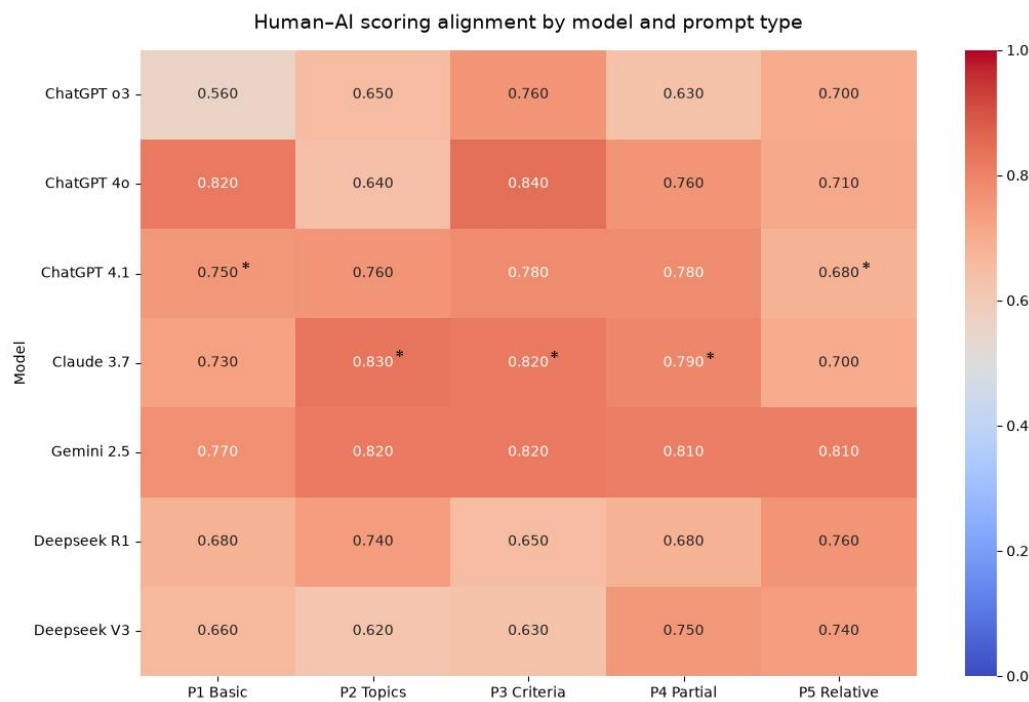


Fig. 2. Visual summary of human-AI alignment by model and prompt type (cell values are correlation coefficients; *indicates no statistically significant difference from the human evaluator, $p > .05$).

5. Conclusion and Suggestions

The present study examined the performance of popular AI systems in scoring open-ended questions using different prompts. The findings indicate that the scoring alignment between instructors and AI models is sensitive to both model selection and prompt content.

On Prompt Sensitivity and Cognitive Loading: The findings confirm that LLMs exhibit high prompt sensitivity; correlation with human experts rose significantly when moving from basic instructions to criteria-based rubrics. This aligns with the ‘Chain-of-Thought’ (CoT) processing theory, suggesting that explicit rubrics act as a cognitive guide for the AI, forcing the model to evaluate knowledge depth and contextual accuracy independently.

Qualitative Analysis of Discrepancies: In cases where models showed ‘generosity bias’ (akin to “leniency error” in psychometrics) scoring higher than the instructor in simpler prompts-qualitative observation suggests that AI models tend to reward the structural volume and formal language of a response. Conversely, the instructor’s lower scores often stemmed from identifying subtle historical inaccuracies that the models failed to categorize as critical errors. A specific instance of this divergence was observed in the responses regarding the Lausanne Conference (Question 3). One student provided a lengthy, structurally coherent essay that received high marks (85/100) from the models, particularly ChatGPT 4o. However, the instructor assigned a significantly lower score (60/100) because the student failed to mention the critical ‘Mosul Question,’ a key learning outcome for that specific lecture. This highlights a ‘hallucination of competence,’ where LLMs tend to conflate linguistic fluency and essay length with factual completeness.

ChatGPT 4.1, for its part, emerged as the sole model showing no statistically significant difference from the human evaluator in unstructured scenarios, specifically the ‘Basic Prompt’ (Prompt 1) and ‘Relative Scoring’ (Prompt 5). While other models like Gemini and Claude optimized their performance with detailed rubrics, ChatGPT 4.1’s alignment with human scoring in basic tasks suggests that its default system prompts or alignment training might favor a more conservative scoring tendency in ambiguous contexts, unlike the ‘generosity bias’ observed in other models.

Specifically, the research found that more recent versions, such as Gemini 2.5 and Claude 3.7, correlate highly with instructors when prompted using well-defined criteria (Prompt 2). Statistical testing for score

differences likewise revealed that the Claude 3.7 model consistently did not reveal statistically significant differences from the instructor across structured prompts (Prompts 2, 3, and 4). This finding suggests that instructor-like question assessment with this model exists, particularly when guided by explicit criteria.

Open-ended question marking has been a topic of widespread debate in the field of educational technology for many decades (see Burrows et al., 2015; Williamson et al., 2006). Following the initial research, rule-based automatic graders indicated a very low measure of reliability as compared to those of human experts (Shermis & Burstein, 2013). However, with recent developments of large language models (LLMs), the systems have been brought nearer to human expert judgments (Milakis et al., 2024; Latif & Zhai, 2024). The findings of our study reveal that the application of prompt-structured model usage is not statistically significantly different from the scores obtained by human graders. This aspect hints at the possibility of successful incorporation of LLMs into education testing procedures. Especially in Prompt 2, where the anticipated responses to the questions were given in their full forms as subject titles, the Claude 3.7 model exhibited both high correlation and no statistical difference in its score and that of the instructor. This observation indicates that the model offers better consistency where there are clear-cut criteria and content headings.

The findings also indicate that AI-human alignment can be very low on more general or open prompts but increases considerably when the prompt is more specific and the criteria for scoring are well-specified. This finding is in line with the literature, which emphasizes that well-specified criteria increase the consistency of evaluation, as indicated in Jonsson and Svingby's (2007) validity principles. The argument advanced by Carlson and Burbano (2025) that multiple and heterogeneous prompt-based analyses are more trustworthy than AI-based tests from a single prompt is supported by the evidence of our research.

Among the main results of the research is that, though there is high correlation with instructor indicated by some models (for instance, Gemini 2.5 and Claude 3.7), the choice of both the model and the prompt emerges as an important element in promoting consistency in evaluation. Previous research identifies discrepancies in performance among various LLMs in the context of educational examinations (Seo et al., 2025) and the potential to obtain disparate outcomes using different prompts (Capdehourat et al., 2025). The findings of the present study align with those of other recent research, including Flodén (2025) and Morjaria et al. (2024). Previous research has shown high agreement among human and AI model judges with prompts that use explicit criteria. Our results once more point to the impact of model and prompt selection on reliable rating, as found in the literature (Seo et al., 2025; Capdehourat et al., 2025).

These results can also be situated within the longer trajectory of automated scoring research. Classic Automated Essay Scoring (AES) systems, from Project Essay Grade (Page, 1966) to e-rater (Attali & Burstein, 2006) and the systems surveyed by Shermis and Burstein (2013) and Burrows et al. (2015), achieved consistency through handcrafted features and large sets of pre-scored training essays. The present findings illustrate how LLMs shift this paradigm: comparable levels of human-AI agreement were reached with no task-specific training data, but only when the prompt itself encoded the scoring construct, effectively replacing the feature-engineering stage of classical AES with rubric-embedded prompt design. Recent studies converge on the same conclusion in different contexts: Yavuz et al. (2025) reported very high reliability ($ICC > .90$) for rubric-based LLM grading of EFL essays, and Farray Rodríguez et al. (2026) showed both that unguided LLMs tend to assign inflated grades and that grounding the model in course-specific content further improves scoring accuracy. The generosity bias observed here under unstructured conditions, and its suppression under criterion-referenced prompts, is therefore consistent with a growing international body of evidence.

The findings of this current study suggest that advanced language models have the potential to produce responses as similar as those of the human evaluators while evaluating open-ended questions. The study also highlights the importance of prompt design as well as the choice of models in maintaining the instructional alignment and accuracy of the assessment process. As a result, in the context of the larger-scale implementation in educational setups, prompts formulation should be treated with the utmost care in order to be inclusive as well as effectively integrated with the assessment criteria, while model performance should be reviewed constantly.

From a pedagogical standpoint, the findings translate into concrete guidance for practice. Instructors who wish to use LLMs in grading open-ended items should (i) treat the AI as a consistent second rater rather than a sole grader, using human-AI disagreements as flags for re-examination; (ii) invest effort in converting their implicit expectations into explicit, weighted rubrics, since this single step produced the largest gains in alignment in this study; (iii) ensure that the factual anchor points of an ideal answer (key events, actors, dates, and cause-effect relations) are explicitly listed in the prompt, because the models otherwise under-penalize factual omissions; and (iv) retain final authority over grades, in line with the human-oversight principle discussed in the Introduction. Used in this way, LLM-assisted scoring can meaningfully reduce the grading workload of instructors teaching large common courses while improving feedback speed and rater consistency. At the policy level, the results support institutional guidelines that mandate rubric-based prompt protocols, disclosure of AI involvement in grading, and systematic human verification of AI-assigned scores, as well as investment in AI-literacy training for faculty (UNESCO, 2023).

Another factor relevant to this study is the fallibility of the human evaluator acting as the standard. Because of differences in evaluation given by different evaluators, it has been proposed by researchers that the use of effective rubrics can eliminate such differences (Moskal and Leydens, 2000; Jonsson and Svingby, 2007). The findings of the analyses indicate that the correlations are moderate to high for the majority of the models under consideration. The statistical significance of scoring differences in most models does not necessarily imply the incorrectness of the AI side. This phenomenon may also be interpreted as the human evaluator committing errors. The high correlation between AI models can be interpreted as an indicator of this situation.

The findings of this study suggest that large language models have the potential to transform evaluation processes in the field of educational technology. However, it is important to note that careful prompt design and the models used remain significant factors in this process.

5.1. Conclusion

The present study investigated the potential for automatic scoring of open-ended questions by entering different prompts into popular artificial intelligence systems. Within the aforementioned scope, a comparative analysis was conducted between the statistical agreement, correlation, and score differences between instructor scores and AI system scores. This comparative analysis was facilitated by the implementation of statistical methodologies, namely the t-test and the Wilcoxon test. The findings indicate that both the AI model utilized and the prompt design exert a substantial influence on the consistency of the scoring.

The most relevant correlation for instructor was found to be largely associated with the Gemini 2.5 and Claude 3.7 models. At the same time, the differences found between the model-generated and the expert scores were found to be statistically non-significant in the case of some categories of the prompts. It was found that the AI models were able to yield scores highly similar to the experts when faced with criterion-defined and specific prompts. By contrast, the level of consistency among the models was found to be rather low in the case of prompts with more generalized or incomplete information. This pattern is, in fact, theoretically expected rather than surprising: because LLMs generate output by predicting the most probable continuation of their input, the more clearly and consistently the scoring construct is specified in the prompt, the more closely the model's output distribution converges on the evaluator's intended criteria, whereas ambiguity in the prompt is reflected directly as variance in the scores.

Overall, the present study demonstrates how artificial intelligence (AI) models are capable of generating scores in conformity with those allocated by humans in the assessment of open-ended questionnaires. To attain this level, however, it is critically important to give priorities to two essential factors, model choice and prompt design. To this, the findings bring forth the need for monitoring AI-driven systems continually, the use of multiple assessment approaches, and the direct involvement of human raters in the assessment process because of the potential for error in these assessments. This, in turn, would contribute to enhancing the effectiveness and reliability of automatic scoring processes in education.

Consequently, for LLMs to serve as valid assessment tools, “prompt engineering” must be re-conceptualized as the rigorous design of scoring protocols. Our findings confirm that validity depends less on the model’s latent intelligence and more on the specificity of the rubric provided, paralleling human rater training principles.

5.2. Limitations

Evaluation of Model Versions and Methodological Consistency The state-of-the-art (SOTA) models available at the time of data collection were utilized in this study. Although LLM technology is advancing rapidly, the study’s core finding that ‘structured prompts reduce hallucinations’ is a generalizable assessment principle that is independent of model architecture.

While the models employed represent the most current versions during the research phase, and LLM performance typically improves with version updates, the positive correlation observed between ‘Prompt Type’ and ‘Scoring Reliability’ is a methodological phenomenon independent of technological shifts. Therefore, the findings offer a valid assessment framework that remains applicable to newer models and future iterations.

Beyond model versioning, several further limitations should be acknowledged explicitly. First, the study is based on a small sample ($N = 24$) drawn from a single course at a single state university; although the composite-score design and the 2,520 item-level scoring events strengthen the internal analysis, the findings should not be generalized to other disciplines, institutions, or student populations without replication. Second, a single instructor served as the human reference standard. While this choice was deliberate, in order to preserve the ecological validity of authentic classroom grading, it means that some human-AI discrepancies may reflect individual rater error rather than model error; future studies should employ at least two independent expert raters and report inter-rater reliability so that human-AI agreement can be benchmarked against human-human agreement.

A further methodological choice should also be stated explicitly. As explained in the Data Analysis section, all inferential analyses were conducted on composite scores rather than on individual items; this choice inevitably leaves the item-level behaviour of the models, including possible differences across question types, outside the scope of the present analysis.

Third, a growing body of research indicates that LLMs can carry linguistic and cultural biases originating from their predominantly English training data and may reproduce contested or one-sided framings of historical events (Yan et al., 2024). Because the present study involved Turkish-language answers on national history, such biases are directly relevant: high statistical agreement with a single instructor does not by itself guarantee cultural or historiographical neutrality, and this should be treated as an explicit limitation. Fourth, the exam questions targeted primarily the knowledge and comprehension levels of Bloom’s taxonomy; the models’ performance on responses requiring analysis, synthesis, and evaluation remains untested. Future research should therefore (i) replicate the design with larger and more diverse samples and with multiple human raters, (ii) extend the assessment to higher-order cognitive levels and to other verbally oriented courses such as literature, (iii) instruct the models to provide written justifications alongside their numerical scores, so that the sources of AI scoring errors (for example, misinterpretations of morphologically inflected historical terms in Turkish) can be analysed directly, (iv) examine retrieval-augmented approaches that ground the model in course materials (Farray Rodríguez et al., 2026), and (v) investigate the longer-term effects of AI-assisted scoring on instructors’ workload and students’ perceptions of fairness.

Declaration on the Use of Artificial Intelligence

The correlation heat maps presented in the Appendices and the summary visual in Figure 2 were produced with the authors’ own Python code, using the matplotlib and seaborn libraries, without the use of any generative AI tool. The flow diagram in Figure 1 was created with the visualization tool drawio.com. No generative AI tool was used in the design of the study, the collection or analysis of the data, or the interpretation of the findings; the AI models listed in the Methodology were themselves the objects of

investigation, not authoring tools. A generative AI tool (Anthropic's Claude) was utilized solely for English language editing and proofreading to improve the academic fluency of the text. The authors reviewed and verified all visual and textual outputs and take full responsibility for the content of the manuscript.

Acknowledgments

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. The authors would like to thank the students who participated in this study.

References

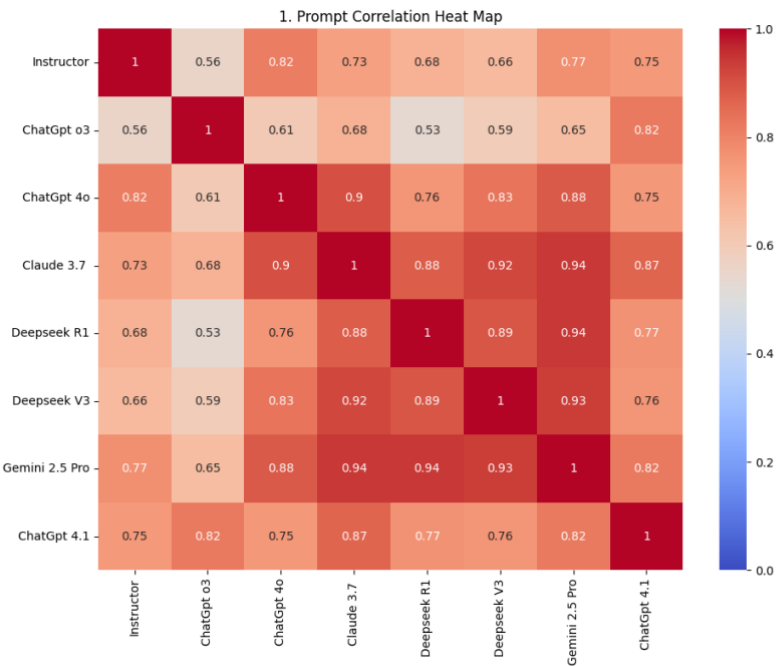
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning and Assessment*, 4(3).
- Badger, E., & Thomas, B. (1992). Open-ended questions in reading. *Practical Assessment, Research & Evaluation*, 3(4), 1991-1993.
- Bektaş, M., & Kudubeş, A. A. (2014). Bir ölçme ve değerlendirme aracı olarak yazılı sınavlar. *Dokuz Eylül Üniversitesi Hemşirelik Fakültesi Elektronik Dergisi*, 7(4), 330-336.
- Brucks, M., & Toubia, O. (2025). Prompt architecture induces methodological artifacts in large language models. *PLoS one*, 20(4), e0319159. <https://doi.org/10.1371/journal.pone.0319159>
- Brown, H. D. (2004). *Language assessment, principles and classroom practices*. USA: Longman.
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1), 60-117. <https://doi.org/10.1007/s40593-014-0026-8>
- Capdehourat, G., Amigo, I., Lorenzo, B., & Trigo, J. (2025). On the effectiveness of LLMs for automatic grading of open-ended questions in Spanish. *arXiv preprint arXiv:2503.18072*. <https://doi.org/10.48550/arXiv.2503.18072>
- Carlson, N., & Burbano, V. (2025). The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*. Advance online publication. <https://doi.org/10.1002/smj.70023>
- Cooney, T. J., Sanchez, W. B., Leatham, K., & Mewborn, D. S. (2004). Open-ended assessment in math: A searchable collection of 450+ questions. "Open-ended Assessment in Math."
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).
- Farray Rodríguez, J., Fernández-García, A. J., & Verdú, E. (2026). Grading open-ended questions using LLMs and RAG. *Expert Systems*, 43(1), e70174. <https://doi.org/10.1111/exsy.70174>
- Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British educational research journal*, 51(1), 201-224. <https://doi.org/10.1002/berj.4069>
- He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). Does Prompt Formatting Have Any Impact on LLM Performance?. *arXiv preprint arXiv:2411.10541*. <https://doi.org/10.48550/arXiv.2411.10541>
- Holmes, W., Iniesto, F., Anastopoulou, S., & Boticario, J. G. (2023). Stakeholder perspectives on the ethics of AI in distance-based higher education. *International Review of Research in Open and Distributed Learning*, 24(2), 96-117. <https://doi.org/10.19173/irrodl.v24i2.6089>

- Jauhiainen, J. S., & Guerra, A. G. (2024). Evaluating Students' Open-ended Written Responses with LLMs: Using the RAG Framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large. *arXiv preprint arXiv:2405.05444*. <https://doi.org/10.48550/arXiv.2405.05444>
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, 2(2), 130-144. <https://doi.org/10.1016/j.edurev.2007.05.002>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons.
- Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, 100210. <https://doi.org/10.1016/j.caeai.2024.100210>
- Linn, R., & Miller, M. (2005). *Measurement and Assessment in Teaching* (9th ed.). Merrill-Prentice Hall.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Milakis, E. D., Argyrakou, C. C., Loukisas, T., Vangeli, D. G., Katsarou, M. C., & Lalou, A. (2024). Comparative analysis of large language models as AI assistants for educational Assessment: A case study in steam education. *International Journal of Computing and Artificial Intelligence*, 5(2), 51-55. <https://doi.org/10.33545/27076571.2024.v5.i2a.97>
- Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical program. *International Medical Education*, 3(1), 32-43. <https://doi.org/10.3390/ime3010004>
- Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical assessment, research, and evaluation*, 7(1).
- OECD. (2023). *OECD digital education outlook 2023: Towards an effective digital education ecosystem*. OECD Publishing.
- Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5), 238-243.
- Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A., & Gama, K. (2023, September). Large language models for education: Grading open-ended questions using chatgpt. In *Proceedings of the XXXVII brazilian symposium on software engineering* (pp. 293-302). <https://doi.org/10.1145/3613372.3614197>
- Popham, W. J. (2020). *Classroom Assessment: What Teachers Need to Know (Ninth Edition)*. University of California, Los Angeles.
- Seo, H., Hwang, T., Jung, J., Kang, H., Namgoong, H., Lee, Y., & Jung, S. (2025). Large Language Models as Evaluators in Education: Verification of Feedback Consistency and Accuracy. *Applied Sciences* (2076-3417), 15(2). <https://doi.org/10.3390/app15020671>
- Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.
- The jamovi project (2024). *jamovi*. (Version 2.6) [Computer Software]. Retrieved from <https://www.jamovi.org>.
- UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO.

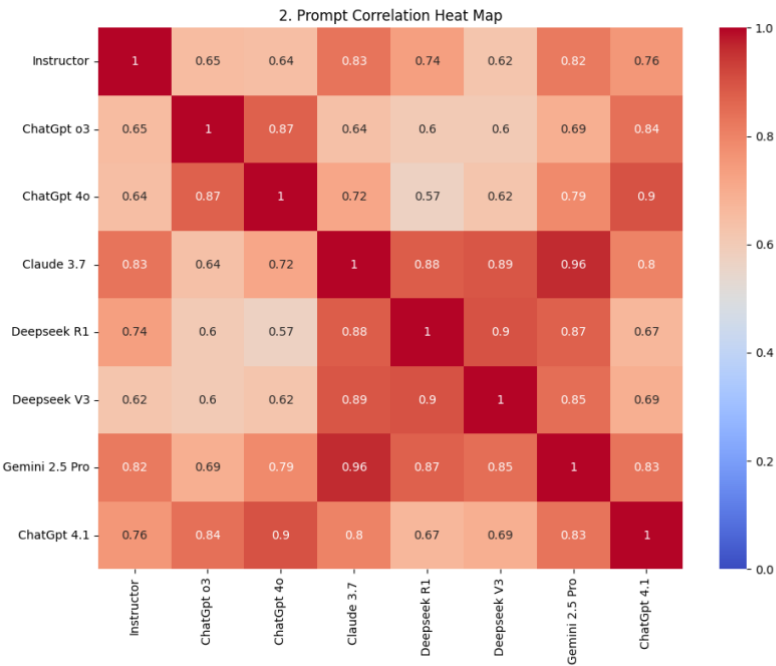
- Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (Eds.). (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah: Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780415963572>
- Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90-112. <https://doi.org/10.1111/bjet.13370>
- Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56(1), 150-166. <https://doi.org/10.1111/bjet.13494>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators?. *International journal of educational technology in higher education*, 16(1), 39. <https://doi.org/10.1186/s41239-019-0171-0>

APPENDIX

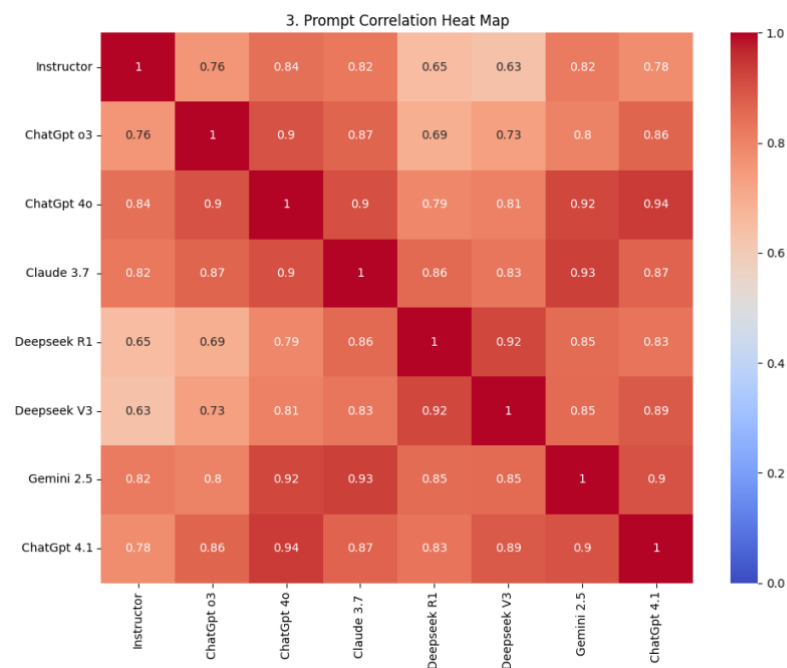
Basic Prompt Correlation Heatmap (A.1)



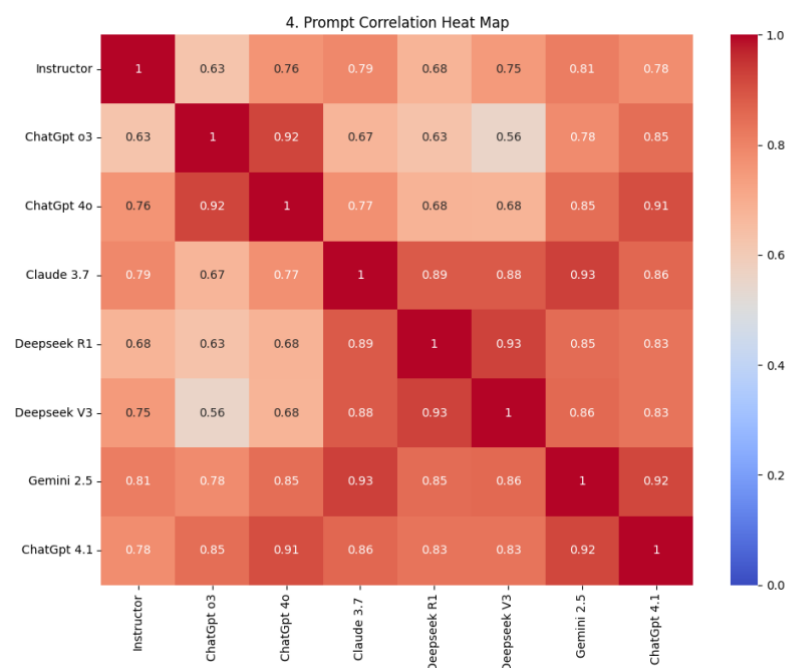
Prompt based on the expected topic headings correlation heatmap (A.2)



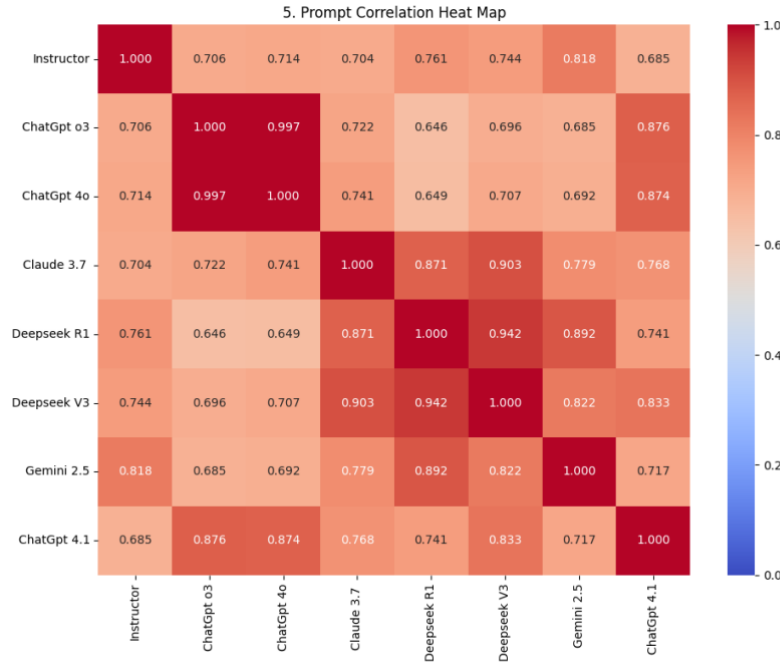
The prompt based on a set of criteria and a percentage correlation heatmap (A.3)



Prompt based on partial coverage and language proficiency correlation heatmap (A.4)



Prompt for relative scoring correlation heatmap (A.5)



Prompt texts used and translations (A.6)

Prompt 1 (Basic Prompt)

Turkish:

Aşağıda sunulan üniversite öğrencisi yanıtlarını, “Atatürk İlkeleri ve İnkılap Tarihi” dersi bağlamında, her bir soru için 0 ile 100 arasında olacak şekilde puanlayınız.

English:

Please grade the following undergraduate student responses, in the context of the course “Atatürk’s Principles and History of the Turkish Revolution”, by assigning a score between 0 and 100 for each question.

Prompt 2 (Prompt based on the expected topic headings)

Turkish:

Her bir öğrenci yanıtını, ilgili soru için aşağıda belirtilen temel konu başlıklarının kapsanma düzeyine göre 0-100 arası puanlayınız.

- **1. Soru:** 1. Dünya Savaşı’nın genel yapısı ve Osmanlı Devleti’nin savaş içindeki durumu
Beklenen başlıklar: I. Dünya Savaşı’nda cepheler, Taarruz cepheleri, Savunma cepheleri, Yardım cepheleri, Savaşın sonu ve ABD, Wilson İlkeleri, Gizli antlaşmalar, Barış/ateşkes anlaşmaları, Mondros sonrası cemiyetler.
- **2. Soru:** Milli Mücadele Dönemi ve Türk İstiklal Harbi
Beklenen başlıklar: Havza/Amasya Genelgesi, Erzurum/Sivas Kongresi, Temsil Heyeti, Misakı Milli, İstanbul’un işgali vb.
- **3. Soru:** Lozan Konferansı’nın tarihsel önemi ve antlaşmanın genel değerlendirmesi
Beklenen başlıklar: Lozan süreci, konferans heyeti, görüşülen başlıklar, temel maddeler, Lozan sonrası meseleler.

English:

Please grade each student response on a scale of 0-100, according to the extent to which the following core topics for each question are addressed.

- **Q1:** General structure of World War I and the Ottoman Empire's role
Expected topics: Fronts in WWI, offensive/defensive/aid fronts, end of the war and the USA, Wilson Principles, secret treaties, peace/armistice agreements, post-Mondros societies.
- **Q2:** Intellectual phase of the National Struggle and Turkish War of Independence
Expected topics: Havza/Amasya Circulars, Erzurum/Sivas Congress, Delegation of Representatives, National Pact, occupation of Istanbul, etc.
- **Q3:** Historical significance of the Lausanne Conference and overall evaluation of the treaty
Expected topics: Lausanne process, delegation, main topics discussed, key provisions, unresolved issues.

Prompt 3 (The prompt based on a set of criteria and a percentage)

Turkish:

Her bir öğrenci yanıtını aşağıdaki değerlendirme kriterleri ve yüzde ağırlıklarına göre değerlendirerek, toplamda 0-100 arasında puanlayınız:

- İçerik doğruluğu (%30)
- Bilgi derinliği (%30)
- Analiz ve yorumlama yeteneği (%10)
- Kronolojik ve bağlamsal anlayış (%10)
- Tutarlılık ve mantıksal akış (%10)
- Dil ve anlatım (%10)

English:

Evaluate each student response according to the following criteria and weighting, and assign a total score out of 100:

- Content accuracy (30%)
- Depth of knowledge (30%)
- Analytical and interpretive ability (10%)
- Chronological and contextual understanding (10%)
- Consistency and logical flow (10%)
- Language and expression (10%)

Prompt 4 (Prompt based on partial coverage and language proficiency)

Turkish:

Bir öğrenci yanıtı, beklenen başlıklardan en az dört tanesini kapsıyorsa ve anlatımı yeterliyse, 0-100 arasında puanlayınız.

English:

If a student's response covers at least four of the expected topics for the question and is adequately expressed, please assign a score between 0 and 100.

Prompt 5 (Prompt for relative scoring:)

Turkish:

Yanıtların tümünü, kendi aralarında karşılaştırmalı olarak, en güçlü yanıtı en yüksek puan alacak şekilde 0-100 arasında göreceli olarak puanlayınız.

English:

Please grade all responses on a relative basis, assigning the highest score to the best answer and lower scores to less comprehensive answers, using a 0-100 scale.

Normality Test Results (A.7)

			Shapiro-Wilk	
		N	W	p
Prompt 1	Instructor	24	0.941	0.172
	ChatGPT o3	24	0.959	0.419
	ChatGPT 4o	24	0.891	0.014
	Claude 3.7	24	0.889	0.013
	Deepseek R1	24	0.896	0.017
	Deepseek V3	24	0.803	<.001
	Gemini 2.5	24	0.845	0.002
	ChatGPT 4.1	24	0.958	0.391
			Shapiro-Wilk	
		N	W	p
Prompt 2	Instructor	24	0.941	0.172
	ChatGPT o3	24	0.942	0.184
	ChatGPT 4o	24	0.971	0.685
	Claude 3.7	24	0.903	0.025
	Deepseek R1	24	0.927	0.082
	Deepseek V3	24	0.921	0.061
	Gemini 2.5	24	0.902	0.023
	ChatGPT 4.1	24	0.952	0.298
			Shapiro-Wilk	
		N	W	p
Prompt 3	Instructor	24	0.941	0.172
	ChatGPT o3	24	0.962	0.478
	ChatGPT 4o	24	0.883	0.009
	Claude 3.7	24	0.913	0.042
	Deepseek R1	24	0.804	<.001
	Deepseek V3	24	0.775	<.001
	Gemini 2.5	24	0.825	<.001
	ChatGPT 4.1	24	0.700	<.001
Prompt 4	Instructor	24	0.941	0.172
	ChatGPT o3	24	0.973	0.737
	ChatGPT 4o	24	0.949	0.263
	Claude 3.7	24	0.881	0.009
	Deepseek R1	24	0.798	<.001
	Deepseek V3	24	0.815	<.001
	Gemini 2.5	24	0.901	0.023
	ChatGPT 4.1	24	0.917	0.051
			Shapiro-Wilk	
		N	W	p
Prompt 5	Instructor	24	0.941	0.172
	ChatGPT o3	24	0.962	0.481
	ChatGPT 4o	24	0.966	0.560
	Claude 3.7	24	0.854	0.003
	Deepseek R1	24	0.922	0.064
	Deepseek V3	24	0.788	<.001
	Gemini 2.5	24	0.923	0.067
	ChatGPT 4.1	24	0.894	0.016