2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #51587

Evaluation of Generative AI Models to Support Student Critical Thinking in Circuit Analysis

Zimeng Guo, The Ohio State University
Ezenwa Opara, Purdue University Fort Wayne
Dr. Bin Chen, Purdue University Fort Wayne

Evaluation of Generative AI Models to Support Student Critical Thinking in Circuit Analysis

Abstract

This research paper presents a statistical comparison of three leading large language models (OpenAI's GPT-5, Google's Gemini 2.5 Pro, and xAI's Grok-4.1) when solving undergraduate circuit analysis problems. The rapid spread of large language models in engineering education requires a systematic empirical assessment of their capabilities, especially in fields that demand complex multimodal synthesis, such as circuit analysis. Using a paired-comparison experimental design with 360 unique problems across 18 chapters and 27 knowledge points, we evaluated model performance across six categories: circuit diagram interpretation, procedural reasoning, formula selection, circuit configuration analysis, computational precision, and plot interpretation. Linear mixed-effects models revealed that statistically, GPT-5 and Gemini 2.5 Pro performed equivalently. Conversely, Grok-4.1 demonstrated marked performance deficits, particularly in scenarios demanding visual interpretation. In AI × image type interactions, GPT-5 possessed a marginal superiority in circuit diagram interpretation, while Gemini had a slight advantage in plot or waveform analysis. The equivalent performance of the two methods was also shown in K-means clustering. These findings provide guidelines for engineering educators. Although GPT-5 and Gemini 2.5 Pro are equally reliable for standard circuit analysis, different tasks should be considered with different selections of models, which will be discussed below.

Introduction

With the wide application of large language models, educators and students are increasingly using them in engineering education [1][2]. Although large language models are very powerful in solving mathematical calculations, problem decomposition, and even complex analysis, their performance in electronic engineering has not been fully explored. Circuit analysis is a particularly challenging field for artificial intelligence (AI) evaluation because it integrates visual-spatial reasoning for circuit topology interpretation, theoretical knowledge application for formula selection, procedural reasoning for systematic problem decomposition, and computational precision for numerical calculations [3]. The complexity of combining text descriptions with chart data makes circuit analysis an ideal platform for evaluation.

For educators and students seeking to effectively utilize AI tools, understanding the relative advantages and limitations of existing models is crucial for informed adoption. Therefore, this study addresses three research questions: (1) Do GPT-5, Gemini 2.5 Pro, and Grok-4.1 differ significantly in overall circuit analysis performance? (2) Are performance differences moderated by problem characteristics, specifically the presence of visual elements? (3) Can we identify practical guidelines for AI tool selection based on specific strengths? To answer these questions, we designed a paired-comparison experiment using 360 circuit analysis problems evaluated across three AI models, followed by a multi-phase statistical analysis combining non-parametric tests, linear mixed-effects modeling, and unsupervised clustering.

Methodology

- Data collection

The dataset of this study was constructed from a comprehensive undergraduate electrical engineering course that covered everything from basic concepts to advanced frequency domain analysis. From each of the 18 chapters, 20 questions were systematically selected, generating 360 unique circuit analysis problems. Problems ranged from basic Ohm's law applications to complex Laplace transform analysis and active filter design. All problems were drawn from a single commercially available textbook, and their use in this study was limited to evaluating AI model outputs rather than reproducing or distributing the textbook content. This approach is consistent with educational fair use guidelines, as the problems served as standardized prompts for AI evaluation rather than as instructional materials.

Each question was independently submitted to OpenAI's GPT-5, Google's Gemini 2.5 Pro, and xAI's Grok-4.1. The standard prompt requires a complete problem-solving solution, including all intermediate steps and the final answer. A total of 1,080 observations were generated during the response generation process (360 questions × 3 AI models). Well-trained student assessors, familiar with circuit analysis standards and solution methods, used structured rules to evaluate each AI response.

- Data preprocessing

Before the statistical analysis, we implemented several data transformation and validation steps to ensure the validity of the analysis. The six evaluation categories employed an ordinal scale of 0 to 4. However, for Categories 1 (circuit diagram interpretation), 4 (circuit configuration analysis), and 6 (plot interpretation), a score of zero indicates that the problem lacks relevant visual content. Therefore, when a problem did not contain a circuit diagram, Category 1 and Category 4 were recoded as "not applicable" (NA) instead of zero.

For convenience of analysis, several compound variables were calculated. The image presence indicator is created as a binary flag (has_circuit_diagram, has_plot, has_any_image). The conditional performance score calculates only the observation results where relevant images exist. These conversions retain the complete dataset while supporting aggregation and hierarchical analysis.

- Evaluation Categories

The multi-dimensional scoring rubric assessed six different performance categories:

Category 1: Ability to read and understand circuit diagrams presented in images.

Category 2: Ability to break down problem-solving procedures into clear, logical steps.

Category 3: Ability to select appropriate formulas and conduct calculations.

Category 4: Ability to analyze circuit configuration and conduct calculations.

Category 5: Accuracy in the use of equations and numerical calculations.

Category 6: Ability to read and understand plots presented in images.

- Statistical analysis

The analytical strategy proceeded through four phases to ensure the robustness of findings through convergent evidence.

Phase 1. Descriptive Analysis: For each category and AI combination, we computed the mean, median, standard deviation, interquartile range, skewness, and kurtosis. Because there are many scores at maximum (ceiling effects), normality was violated for all categories, so we selected non-parametric tests for basic inference.

Phase 2. Paired Comparisons: We used Wilcoxon signed-rank tests for pairwise model comparisons within each category, with Benjamini–Hochberg false discovery rate correction [4] applied to control Type I errors across the six category tests. The Wilcoxon signed-rank test is a non-parametric alternative to the paired t-test that does not assume normally distributed differences; instead, it ranks the absolute differences between paired observations and compares the sum of positive and negative ranks [5]. The Benjamini–Hochberg procedure controls the expected proportion of false positives (false discovery rate) when conducting multiple hypothesis tests, offering greater statistical power than traditional family-wise corrections such as Bonferroni.

Phase 3. Linear Mixed-Effects Modeling: We employed mixed-effects models (LMMs), which extend standard regression by simultaneously estimating fixed effects (the systematic influence of predictors such as AI model and category) and random effects (variability attributable to grouping structures such as individual problems and topics). This framework is well suited to repeated-measures designs in which each problem is scored by multiple AI models, because it accounts for within-problem correlation and between-topic heterogeneity without requiring independence among observations [6]. The model specification was as follows:

$$\text{Score\_ij} = \beta_0 + \beta_1(\text{AI}) + \beta_2(\text{Category}) + \beta_3(\text{has\_circuit\_diagram}) + \beta_4(\text{has\_plot}) + \beta_5(\text{AI} \times \text{Category}) + \beta_6(\text{AI} \times \text{has\_circuit\_diagram}) + \beta_7(\text{AI} \times \text{has\_plot}) + \beta_8(\text{Category} \times \text{has\_circuit\_diagram}) + \beta_9(\text{Category} \times \text{has\_plot}) + \text{u\_problem} + \text{u\_topic} + \varepsilon_{ij}$$

where *AI*, *category*, *image* indicate presence, and all two-way interactions capture systematic influences, *u_problem* captures repeated-measures structure, and *u_topic* captures clustering. Chapters within Knowledge_Point were included where supported by variance structure. Models were estimated using restricted maximum likelihood with the BOBYQA optimizer in the lme4 package [7]. Post-hoc comparisons used estimated marginal means with Tukey's adjustment for multiplicity. Model diagnostics included residual plots, Q-Q plots, and variance inflation factors for multicollinearity assessment.

Phase 4. Pattern Discovery: To identify latent structures in the performance data, we applied K-means clustering to problem-level features. The optimal number of clusters was determined using the elbow method and silhouette analysis. K-means is an unsupervised learning algorithm that partitions observations into k groups by iteratively minimizing the within-cluster sum of squared distances to each cluster centroid [8]. The

elbow method plots the total within-cluster variance against the number of clusters and identifies the point at which additional clusters yield diminishing reductions in variance. Silhouette analysis complements this by measuring how similar each observation is to its own cluster compared to neighboring clusters, with values closer to 1.0 indicating well-separated groupings [9].

Having established the analytical framework, we now present the findings from each phase. The results are organized by analysis type, beginning with descriptive statistics, proceeding through inferential comparisons, and concluding with cluster-based pattern discovery.

Results

- Descriptive statistics

Table 1 presents the descriptive statistics for each category. First, for all text-based questions, Categories 2, 3, and 5 show uniformly high performance across all models. Means range from 3.14 to 3.92, and strong negative skewness indicates ceiling effects. Second, for circuit diagram questions, Categories 1 and 4 show intermediate means due to structural zeros, with approximately 36% of problems lacking circuit diagrams. Third, for plot interpretation, Category 6 shows the lowest means, because only 6.7% of problems contain plots.

Table 1. Descriptive Statistics by Category and AI Model

Category	AI Model	N	Mean	SD	Median	Skewness	% Nonzero
Cat1	GPT-5	360	2.51	1.92	4.00	-0.53	63.9%
	Gemini 2.5 Pro	360	2.46	1.90	4.00	-0.48	63.9%
	Grok-4.1	360	2.01	1.75	2.00	0.02	64.7%
Cat2	GPT-5	360	3.86	0.48	4.00	-3.67	100%
	Gemini 2.5 Pro	360	3.84	0.52	4.00	-3.74	99.7%
	Grok-4.1	360	3.33	1.03	4.00	-1.12	100%
Cat3	GPT-5	360	3.92	0.38	4.00	-5.27	100%
	Gemini 2.5 Pro	360	3.92	0.41	4.00	-5.76	99.7%
	Grok-4.1	360	3.34	1.03	4.00	-1.16	100%
Cat5	GPT-5	360	3.73	0.62	4.00	-2.20	100%
	Gemini 2.5 Pro	360	3.68	0.66	4.00	-2.13	99.7%
	Grok-4.1	360	3.14	1.11	4.00	-0.78	100%

Note: Categories 4 and 6 omitted for reasons of space; patterns parallel Categories 1 and 6 respectively. Nonzero percentages for Cat1/Cat4/Cat6 indicate problems containing relevant images.

- Mixed-effects model results

The mixed-effects model explained 84.7% of variance, and the combined fixed and random effects explained 91.9% of variance. Table 2 presents the estimated marginal means for each AI model, averaged across categories. GPT-5 achieved an estimated marginal mean of 2.79 (SE = 0.084), Gemini 2.5 Pro achieved 2.81 (SE = 0.084), and Grok-4.1 achieved 2.32 (SE = 0.084). Pairwise comparisons revealed no significant difference between GPT-5 and Gemini, but both significantly outperformed Grok-4.1.

Table 2. Pairwise Model Comparisons from the Mixed-Effects Model

Comparison	Estimate	SE	z-ratio	p-value	Cohen's d
Gemini – GPT-5	0.020	0.032	0.616	.812	0.02
Gemini – Grok-4.1	0.489	0.032	15.079	<.0001***	0.49
GPT-5 – Grok-4.1	0.469	0.032	14.470	<.0001***	0.47

Note: P-values adjusted using Tukey's method for comparing three estimates. *** p <.001

- Interaction effects

The coefficient for the Grok-4.1 × has_circuit_diagram interaction was highly statistically significant, indicating that Grok-4.1's performance deficit was substantially larger on problems containing circuit diagrams. Neither GPT-5 nor Gemini showed significant circuit diagram interactions, suggesting that these models handle visual content more robustly.

To clarify this interaction, Table 3 presents performance stratified by image type. For problems containing circuit diagrams, GPT-5 demonstrated a confidence index of 74.8%, compared to Gemini's 65.7%. For text-only problems, the models performed identically. For plot or waveform problems, Gemini showed a slight advantage, although this did not reach significance due to the limited sample size.

Table 3. Performance Stratified by Image Type

Image Type	AI Model	N	Text Score	Conf. Index	95% CI
Circuit Diagram	GPT-5	230	3.78	74.8%	69.2-80.4%
	Gemini 2.5 Pro	230	3.75	65.7%	59.5-71.8%
	Grok-4.1	233	2.99	41.2%	34.9-47.5%
No Image	GPT-5	107	3.96	90.7%	85.1-96.2%
	Gemini 2.5 Pro	107	3.93	90.7%	85.1-96.2%
	Grok-4.1	103	3.89	85.4%	78.6-92.2%
Plot/Waveform	GPT-5	23	3.83	82.6%	67.1-98.1%
	Gemini 2.5 Pro	23	3.96	91.3%	79.8-100%
	Grok-4.1	24	3.39	66.7%	47.8-85.6%

Note: Text Score = mean of Categories 2, 3, 5. Confidence Index = proportion achieving maximum scores on all applicable categories.

- Knowledge point analysis

To identify topic-specific performance patterns, we conducted a stratified analysis across all 27 knowledge points. Table 4 presents the text-only composite scores, which are computed from the mean of Categories 2, 3, and 5. Among the 27 knowledge points analyzed, the majority showed comparable performance between GPT-5 and Gemini. However, the four knowledge points exhibited notable differentials.

Table 4. Performance by Knowledge Point (Selected Topics with Largest Differentials)

Knowledge Point	N	GPT-5	Gemini	Grok	Δ (G-Ge)	Dominant
Kirchhoff's Laws	14	3.93	3.33	2.19	+0.60*	GPT-5
Response of First-Order RL/RC	20	3.75	3.63	2.00	+0.12	GPT-5
Fourier Series	20	3.80	3.95	3.37	-0.15	Gemini
The Difference-Amplifier Circuit	4	3.33	3.67	3.00	-0.33	Gemini

Note: Δ (G-Ge) = GPT-5 minus Gemini difference. *p <.05 based on paired t-test.

The most striking finding is Kirchhoff's Laws, where GPT-5 achieved a mean score of 3.93 compared to Gemini's 3.33. This finding is notable because the correct application of Kirchhoff's Laws is a prerequisite for most circuit diagram problems, potentially explaining GPT-5's overall advantage in visual-spatial reasoning tasks.

GPT-5 also demonstrated an advantage in the Response of First-Order RL and RC Circuits, suggesting better handling of time-domain differential equations and transient analysis. Conversely, Gemini showed advantages in the Fourier Series and the Difference-Amplifier Circuit, indicating potential strengths in frequency domain analysis and operational amplifier configurations. However, aside from Kirchhoff's Laws, these differences did not reach statistical significance.

- Cluster analysis

K-means clustering with k = 4 identified four distinct problem profiles, as shown in Table 5. Cluster 1 represented problems where Grok-4.1 showed severe performance collapse, particularly on circuit diagram interpretation; GPT-5 and Gemini performed equivalently in this cluster, while Grok-4.1 averaged only 1.65. Clusters 2 and 3 showed low variance across AI systems. Cluster 2 contained high-performance problems in which all models succeeded, while Cluster 3 contained challenging problems in which all models struggled. Cluster 4 showed moderate variance with mixed difficulty patterns.

Table 5. Problem Cluster Characteristics

Cluster	N	Avg Score	AI Variance	Interpretation
1	88	2.58	0.907	Grok-4.1 Collapse (circuit diagrams)
2	118	3.25	0.079	High Performance, Low Variance (all models succeed)

3	103	1.97	0.046	Low Performance, Low Variance (all models struggle)
4	23	2.47	0.284	Mixed Difficulty (moderate variance)

Taken together, these results reveal a consistent pattern: GPT-5 and Gemini 2.5 Pro perform at comparable levels overall, while Grok-4.1 lags behind, especially when visual interpretation is required. The following discussion contextualizes these findings and derives practical recommendations for engineering educators.

Discussion

The primary finding is that GPT-5 and Gemini 2.5 Pro demonstrate statistically equivalent overall performance across all evaluation categories. For engineering educators, this finding suggests that either model represents an equally reliable tool for circuit analysis assistance.

Grok-4.1 demonstrated significantly inferior performance compared to both GPT-5 and Gemini 2.5 Pro, with medium effect size deficits in text-based categories and small deficits in image-based categories. The interaction analysis revealed that Grok-4.1's weakness is particularly pronounced for circuit diagram problems, which represents a substantive practical limitation for engineering applications.

Although GPT-5 and Gemini 2.5 Pro demonstrate considerable comprehensive performance, they show complementary advantages in terms of question types. GPT-5 shows statistically significant advantages in circuit diagram interpretation, indicating superior visual-spatial reasoning ability in schematic analysis. In contrast, Gemini 2.5 Pro shows insignificant advantages in terms of plots or waveforms, indicating potential specialization in data visualization. These findings point to an optimal choice strategy, namely, to be more inclined toward GPT-5 in issues emphasizing circuit topology, and to consider Gemini 2.5 Pro in issues involving the interpretation of graphic data.

- Practical implications

1. For general circuit analysis assistance: GPT-5 and Gemini 2.5 Pro are equally reliable choices.

2. For problems with circuit diagrams: GPT-5 demonstrates superior performance. It also provides more reliable outputs in topology extraction, component identification, or Kirchhoff's law application.

3. For problems with plots or waveforms: Gemini 2.5 Pro shows a modest advantage in interpreting graphical data such as frequency response curves, transient waveforms, and Bode plots.

4. Avoid relying on Grok-4.1: Until substantial improvements are demonstrated, Grok-4.1 should not be recommended for circuit analysis coursework.

5. Verify AI outputs on challenging problems: The cluster analysis revealed that 28.6% of problems were challenging for all models. Educators should train students to critically evaluate AI outputs, particularly for complex problems or unfamiliar topic areas.

- Limitations

First, the single-rater design, while ensuring consistency, is open to bias. Future studies should employ multiple trained evaluators with reliability analyses. Second, the sample was drawn from a single textbook and institution, limiting generalizability to other curricula or problem formats. Third, AI models undergo continuous updates, which may not persist across versions. Finally, the study evaluates correctness, but not pedagogical utility. Whether AI explanations enhance student learning remains an open question.

Conclusion

This comparative analysis of GPT-5, Gemini 2.5 Pro, and Grok-4.1 on undergraduate circuit analysis problems provides evidence for model selection in engineering education. The statistical framework revealed that GPT-5 and Gemini 2.5 Pro are functionally equivalent for most applications, with advantages in visual processing. Grok-4.1 demonstrated consistent performance deficits, particularly in multimodal tasks. For educators seeking to integrate AI tools responsibly, these findings offer an empirical foundation for informed adoption decisions.

References

[1] J. Qadir, "Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, 2023, pp. 1–9.

[2] D. Baidoo-Anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.

[3] C. Leon, J. Lipuma, and X. Oviedo-Torres, "Artificial intelligence in STEM education: a transdisciplinary framework for engagement and innovation," *Front. Educ.*, vol. 10, Art. no. 1619888, Jul. 2025, doi: 10.3389/feduc.2025.1619888.

[4] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[5] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[6] H. Singmann and D. Kellen, "An introduction to mixed models for experimental psychology," in *New Methods in Cognitive Psychology*, D. H. Spieler and E. Schumacher, Eds. New York, NY, USA: Routledge, 2019, pp. 4–31.

[7] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[8] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. Roy. Statist. Soc. Ser. C (Appl. Statist.)*, vol. 28, no. 1, pp. 100–108, 1979.

[9] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.