2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC
Host Sponsor UNC Charlotte
ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION
Paper ID #51628

Systematic and Quantitative Evaluation of Generative AI Models for Electrical Circuit Analysis

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

Systematic and Quantitative Evaluation of Generative AI for Electrical Circuit Analysis

Abstract

The recent advancement in generative artificial intelligence (AI) has transformed the way students and educators approach education by personalizing learning experiences, creating dynamic content, and offering efficient administrative support to educators. While previous research has surveyed student perceptions of AI tools like ChatGPT across science and engineering courses and studied student experiences using AI tools for writing assignments through comparative analysis, direct assessment of AI problem-solving accuracy in foundational engineering courses remains underexplored.

We investigated the reliability of three state-of-the-art AI models—Google's Gemini 2.5 Pro, OpenAI's GPT-5, and xAI's Grok 4.1 in solving electrical circuit problems from Electric Circuits, 12th Edition by Nilsson and Riedel. A total of 360 questions were systematically selected from 18 chapters of the book with 20 questions per chapter. The questions consisted of text-based questions and image-based questions containing circuit diagrams and plots. AI responses were evaluated using a six-dimensional analytic rubric on a 4-point scale (1 = Poor, 2 = Fair, 3 = Good, 4 = Excellent; 0 = N/A when diagrams or plots were not present). The six dimensions were: (1) diagram comprehension, (2) stepwise problem decomposition, (3) formula selection and calculation, (4) circuit configuration analysis, (5) equation use and numerical accuracy, and (6) plot interpretation.

The results showed that GPT-5 and Gemini 2.5 Pro performed at nearly the same level (mean scores: 3.85 vs. 3.83, $p = 0.48$), both demonstrated excellent performance with low variability. Grok 4.1 exhibited relatively lower performance (mean: 3.28, $p < 0.001$) with higher inconsistency. This study provides quantitative evaluations showing that AI models have achieved substantial competency in electrical engineering circuits problem-solving, while also highlighting their limitations and potentials in engineering education.

1. Introduction

Generative artificial intelligence (genAI) is changing engineering education by personalizing learning, creating learning materials, and improving teaching efficiency. When major companies began promoting their most advanced AI models as having "PhD-level expertise in any topic," [1] these claims generated both enthusiasm and skepticism among engineering educators. Despite a rapidly growing publication on AI in STEM education, much of the study remains either perception based like surveys [2], [3], [4] or limited to ad hoc question sets. There is a need for systematic, quantitative evaluation of AI performance on discipline-specific problems, especially in engineering education.

1.1 Related Work

Several studies have investigated AI performance in technical education with varying evaluation approaches. Research on programming and physics courses provided quantitative performance metrics[5], [6], but these studies focused mainly on correct/incorrect outcomes rather than detailed assessment of solution methodology or reasoning processes.

A cybersecurity education study [7] employed a weighted efficacy framework (Technical Accuracy, Usefulness, Relevance, and Critical Thinking) to evaluate ChatGPT-4 and Gemini Pro across over 100 questions. Their results showed that ChatGPT outperformed Gemini in code generation (96% vs 76%) and threat detection tasks. However, their evaluation remained task-outcome focused rather than examining the different aspects of problem-solving common in engineering such as procedural clarity, formula selection, circuit analysis, and calculation accuracy.

Another study evaluated the ChatGPT, Gemini, Claude, and Meta AI as tutors specifically to uncover how these large language models can contribute to teaching and learning of general computer science and engineering courses. Their method involved asking models a popular debate relating to computer architecture ("*how is cisc better than risc")* and analyzing the responses based on accuracy, persuasiveness, and depth of explanation. ChatGPT scored 90%, Meta AI 85.71%, Claude 77.78%, and Gemini 71.42% for accuracy [8].

Also, a high school mathematics study [9] evaluated GPT-4o and Gemini 1.5 on 60 questions across seven mathematical branches (grades 7-12), comparing unimodal (image-only) and multimodal (text-image) inputs while testing robustness to blur and rotation artifacts, finding GPT-4o achieved high accuracy in both modalities while Gemini 1.5 performed better with unimodal inputs, with both models struggling particularly in geometry (50% accuracy); however, this assessment similarly focused on answer correctness rather than the granular, multi-dimensional evaluation of solution quality across distinct competency categories employed in the present work.

1.2 Significance and Objectives

Although the existing studies have provided insights into the capabilities of AI across various domains, there still exists a gap in understanding how these models perform across the range of skills needed in engineering problem-solving. Engineering educators do not evaluate only correct final answers; skills like sound procedural reasoning, appropriate formula selection, accurate circuit interpretation are evaluated as well. The significance of this study lies in its comprehensive evaluation framework that mirrors how real engineering assessment is done. We went beyond binary correctness, by assessing six distinct competency categories.

The goal of this study is to provide engineering educators and students with empirical evidence regarding the capabilities of AI models and their limitations. This will enable them to make informed decisions about AI integration in curriculum design, assessment strategies, instructional support, and personal use.

2. Methods

2.1 Data Preparation and Description

We used stratified sampling to select the questions for the study to ensure a reasonable representation of different question types across all chapters; 20 questions were selected from 18 chapters of *Electric Circuits, 12th Edition by Nilsson and Riedel* giving a total of 360 questions. Most chapters consist of a mix of conceptual questions, analytical computation problems, problems with diagram interpretation, problems containing plots or time-domain graphs.

Text-based problems were copied verbatim from the textbook and placed in a text file: No extra information was added; questions were single part. Single problems were chosen from multipart questions and no prompt was included. Files containing questions were named to have consistent naming method such as Chapter_1_Question_2.txt.

Image-based problems included several pieces of information. Some of the images contained only text while others had text with figures such as circuit diagrams, plots or tables. The images were processed by cropping select-questions from the textbook and saving them on the local disk with consistent naming methods such as Chapter_6_Question_8.png

2.2 AI Model Selection and Input Delivery

Three state-of-the-art AI models were selected for this study which include OpenAI's GPT-5, Google's Gemini 2.5 Pro and xAI's Grok 4.1. The generative AI models were selected for this research based on their reported technical capabilities and public accessibility. OpenAI claimed that GPT-5 has PhD level expertise with strong computational reasoning setting new state-of-the-art across Math by achieving 94.6% score on AIME 2025[1]. Gemini 2.5 Pro is capable of advanced reasoning and solving complex problems [10] while Grok 4.1 is said to be more like human with strong reasoning ability [11].

We used a Python script to interact with the OpenAI API for GPT-5. Text was read from text files and sent as UTF-8 encoded data while image files were encoded as base64 and submitted via API. Gemini 2.5 Pro and Grok 4.1 were accessed via the web interface with manual submission. Text from text files were copied and pasted in the chat window and images were uploaded and accompanied with the prompt to solve the problem step by step.

To maintain neutrality for both API and web interface access, each question was submitted only once (unless timeout), no follow up correction or additional prompting and the instruction remained "*Please solve the following problem step by step.*" This simulates how a student would prompt generative AI models to provide solutions.

2.3 Evaluation Framework

We developed a systematic rubric to evaluate the model response. The rubric consists of six performance categories (Table 1) which reflect essential engineering problem-solving skills. Each category was scored on a 5-point scale (Table 2), with not all categories necessarily applying to every question.

Table 1 Evaluation rubric categories

S/N	Category	Description
1	Diagram interpretation	Ability to read and understand circuit diagrams presented in images
2	Problem solving steps	Ability to break down problem-solving procedures into clear, logical steps
3	Formula selection	Ability to select appropriate formulas and conduct calculations
4	Circuit analysis	Ability to analyze circuit configuration and conduct calculations
5	Mathematical accuracy	Accuracy in the use of equations and numerical calculations
6	Plot interpretation	Ability to read and understand plots presented in images

Table 2 Rating Score Description

Score	Meaning
0	Not applicable
1	Poor
2	Fair
3	Good
4	Excellent

2.4 Data Analysis Methods

The data analysis methods ranged from basic descriptive statistics to advanced statistical testing and multidimensional aggregation to comprehensively evaluate model performance.

a. Statistical Calculations

We calculated the descriptive statistics after excluding N/A values (which were scored as 0 when categories did not apply to a particular question) to ensure accurate statistical representation.

b. Statistical Tests

We used analysis of variance (ANOVA) F- test to determine if there are significant differences between models, followed by pairwise t-tests for independence between models and p-values for statistical significance testing.

c. Performance Analysis

We conducted extensive performance analysis at multiple levels which include chapter-by-chapter aggregation and analysis, model-by-model performance, category-by-category investigation across 6 evaluation dimensions.

d. Overall Score Calculation

For each response, the mean score was computed across all six categories:

$$\text{OverallScore}_i = \frac{1}{6}\sum_{j=1}^{6} \text{Category}_{ij}$$

where $\text{Category}_{ij} \in \{1,2,3,4,\text{NaN}\}$ and NaN values were excluded.

e. Comparative Analysis

Metrics across all questions per model and per chapter-model combination were aggregated. We identified the worst performing chapters and group categories into visual comprehension (categories 1 and 6) and analytical skills (categories 2-5) to assess model strengths.

f. Trend Analysis

Trend analysis examined performance patterns across the 18 sequential chapters, calculating summary statistics for each chapter to identify difficulty progression and topic-specific challenges.

3. Results

3.1 Overall Model Performance

The overall model performance across all 360 questions shows that GPT-5 came as top performer by mean score (Table 3), but the margin over Gemini 2.5 Pro was very small ($\Delta \approx 0.02$). Both models showed excellent consistency with low variability (SD ≈ 0.41 – 0.42). On the other hand, Grok 4.1 showed a significantly lower performance with high variability (SD =1.02), indicating inconsistent reliability. All models achieved median scores of 4.0, showing clustering toward the excellent performance.

Table 3 Mean Scores Across All 360 Questions

Model	Mean Score	Standard Deviation	Coefficient of Variation	Performance Notes
GPT-5	3.85	0.41	10.65%	Excellent
Gemini 2.5 Pro	3.83	0.42	11.00%	Excellent
Grok 4.1	3.28	1.02	31.10%	Good (highly variable)

3.2 Statistical Significance Testing

ANOVA testing revealed statistically significant difference between models overall ($F = 80.68$, $p < 0.0001$). However, pairwise comparison showed that GPT-5 and Gemini 2.5 Pro had close performance, while both significantly outperformed Grok 4.1 (Table 4).

Table 4 Pairwise Model Comparison

Comparison	Significance Level	p-value
GPT-5 vs Gemini 2.5 Pro	No significant difference	0.48
GPT-5 vs Grok 4.1	Highly significant	< 0.0001
Gemini 2.5 Pro vs Grok 4.1	Highly significant	< 0.0001

3.3 Performance by Category

Analysis of the six evaluation categories showed distinct patterns in model strengths and weaknesses (Table 5). GPT-5 excelled at diagram interpretation, while Gemini 2.5 Pro showed strength in formula selection and plot reading. Both models demonstrated relative weakness in circuit analysis and calculation accuracy. Grok 4.1 consistently underperformed across all categories, with circuit analysis representing its most significant weakness.

Table 5 Performance by Evaluation Category

Category	GPT-5	Gemini 2.5 Pro	Grok 4.1
1. Circuit Diagram Reading	**3.93**	3.85	3.10
2. Procedural Clarity	3.86	3.86	3.33
3. Formula Selection	3.92	**3.93**	3.34
4. Circuit Analysis	3.77	3.68	2.97
5. Calculation Accuracy	3.73	3.69	3.14
6. Plot Reading	3.79	**3.92**	3.38
Strongest Category	Diagram (3.93)	Formula (3.93)	Plot (3.38)
Weakest Category	Calculation (3.73)	Circuit Analysis (3.68)	Circuit Analysis (2.97)

Visual vs Analytical Performance: Categories were grouped into visual comprehension (diagram and plot interpretation) and analytical skills (procedural clarity, formula selection, circuit analysis, calculation accuracy). Interesting patterns emerged as Gemini 2.5 Pro showed stronger visual comprehension (3.89) compared to its analytical performance (3.79). On the other hand, GPT-5 showed the most balanced profile with nearly identical visual (3.86) and analytical (3.82) capabilities. Grok 4.1 showed slight strength in visual tasks (3.24) over analytical tasks (3.20), though both remained substantially below the other models.

3.4 Performance by Chapter

- Chapter-by-chapter trends

The model performance varied significantly across chapters and topics. There was a low performance on foundational topics (Chapters 2-4, 7) especially for Grok 4.1 (Figure 1) but excellent performance across all models for advanced mathematical topics (11-15, 17).

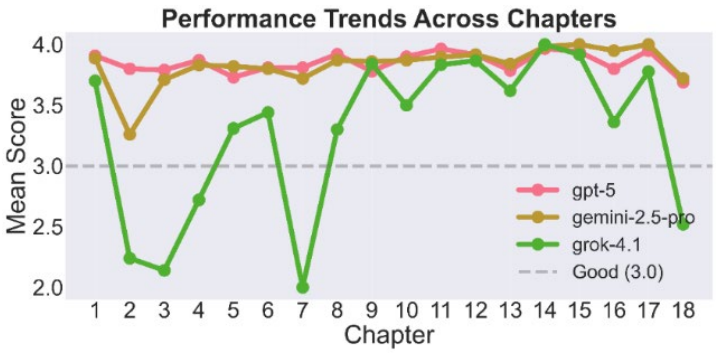

Figure 1 Model Performance Trends Across 18 Chapters. Grok 4.1 demonstrates severe struggles in early chapters (2, 3, 7) with substantial improvement in later chapters. Gemini 2.5 Pro shows early variability that stabilizes. GPT-5 maintains consistent performance throughout

- Best and Worst Performing Topics

The best and worst performing chapters for each model are shown in Table 6 and Table 7 respectively. This highlights the performance of the models on advanced topics with Gemini 2.5 Pro occupying more spots than other models. The worst performance table reveals Grok 4.1's severe struggle with foundational topics.

Table 6 Top 5 Chapter-Model Combinations

Chapter	Model	Mean	Std	Topic
15	Gemini 2.5 Pro	4.00	0.00	Active Filter Circuits
17	Gemini 2.5 Pro	4.00	0.00	The Fourier Transform
14	Grok 4.1	4.00	0.00	Introduction to Frequency-Selective Circuits
14	Gemini 2.5 Pro	3.98	0.08	Introduction to Frequency-Selective Circuits
11	GPT-5	3.96	0.11	Balanced Three-Phase Circuits

Table 7 Bottom 5 Chapter-Model Combinations

Chapter	Model	Mean	Std	Topic
7	Grok 4.1	2.00	0.56	Response of First-Order RL and RC Circuits
3	Grok 4.1	2.14	1.34	Resistors in Series and Parallel
2	Grok 4.1	2.24	1.40	Electrical Resistance (Ohm's Law)
18	Grok 4.1	2.52	0.91	Two-Port Circuits
4	Grok 4.1	2.72	0.85	Node-Voltage Method

- Performance by Topic Category

When aggregating chapters into five topic categories (Table 8), the counterintuitive pattern becomes clear: all models performed best on mathematically complex transforms (Chapters 11–15) and worst on foundational concepts (Chapters 1–4). This suggests that rule-based, systematic problems are easier for AI than conceptual fundamentals requiring physical intuition.

Table 8 Performance by Topic Category

Topic Category	Chapters	GPT-5	Gemini 2.5 Pro	Grok 4.1
Foundational Concepts	1–4	3.84	3.70	2.68

Topic Category	Chapters	GPT-5	Gemini 2.5 Pro	Grok 4.1
Component Methods	5–8	3.82	3.80	3.01
Steady-State Analysis	9–10	3.84	3.87	3.67
Transforms	**11–15**	**3.91**	**3.93**	**3.85**
Advanced Analysis	16–18	3.81	3.89	3.22

3.5 Score Distribution Analysis

The distribution rating across the 4-point scale shows important differences in the model reliability (Table 9). GPT-5 and Gemini 2.5 Pro achieved "Excellent" ratings 87-89% of time with their "Poor" ratings <1%. Grok 4.1 showed a bimodal distribution with 61.8% "Excellent" ratings and 9.9% "Poor" ratings, indicating high unpredictability. This inconsistency makes Grok 4.1 unreliable even though it achieved perfect score in chapter 14 (Table 6).

Table 9 Score Distribution Across Rating Levels

Rating	GPT-5	Gemini 2.5 Pro	Grok 4.1
Excellent (4)	89.4%	87.4%	61.8%
Good (3)	5.6%	6.4%	6.5%
Fair (2)	4.3%	5.8%	21.8%
Poor (1)	0.6%	0.3%	9.9%

4. Discussions

This systematic evaluation of three state-of-the-art AI models on electrical circuit analysis revealed important patterns; the results showed that GPT-5 and Gemini 2.5 Pro demonstrated statistically equivalent performance (means: 3.85 vs 3.83, $p=0.48$) with high consistency (CV ~11%) and deliver excellent ratings 87-89% of the time, while Grok 4.1 performs significantly lower (mean: 3.28, $p<0.001$) with nearly triple the variability (CV=31.1%). Counterintuitively, all models performed better on advanced topics like mathematical transforms (mean: 3.90) than foundational concepts like Ohm's Law (mean: 3.10), suggesting that AI excels at algorithmic, rule-based problems but struggles with physical intuition and conceptual reasoning. Circuit analysis category is the universal weakness across all models, while formula selection represents a collective strength. The models showed strong visual comprehension, especially for Gemini 2.5 Pro, but this does not translate to high configuration analysis. This suggests that the AI models can "see" the diagrams without fully understanding circuit behavior.

The consistent weakness in circuit analysis across all models suggests limitations in spatial reasoning and understanding of electrical circuits. While the models can read the diagrams visually and select the right formula, they struggle to integrate these capabilities into a holistic circuit configuration understanding. Gemini 2.5 Pro consistently gave longer explanation to the problems compared to GPT-5 that concisely went straight to the solution. This long explanation sometimes caused Gemini 2.5 Pro to not perform well on circuit analysis. The long explanation could benefit students/educators who have an idea of the topic, but caution should be applied. Grok 4.1's bimodal performance indicates fundamental architectural or training issues that make

it unpredictable. During its reasoning process and searching through the web, it sometimes reads the exact problem solution on websites like Chegg but totally ignores it until producing a response (sometimes wrong).

5. Limitations

The study used only electric circuit analysis from one textbook (Nilsson & Riedel). Findings may not generalize to other engineering disciplines or textbooks. With numerous genAI out there, the study was limited to GPT-5, Gemini 2.5 Pro, and Grok 4.1. Claude, Copilot, Llama, and Mistral were not included, and results represent a snapshot of the tested set. Also, each question was submitted once. Iterative prompting, chain-of-thought priming, or few-shot examples might significantly improve scores for all models.

Gemini and Grok were accessed via web interfaces while GPT-5 used the API. Interface differences may introduce variability not attributable to model quality alone. And despite a defined rubric, human scoring introduces potential inter-rater variability. A single rater was used; no inter-rater reliability coefficient was computed.

AI models update frequently. Grok 4.1 and Gemini 2.5 Pro may perform differently in current or future versions. Results reflect the 2025 testing period.

6. Conclusions and Recommendations

GPT-5 and Gemini 2.5 Pro are reliable for tasks involving formula selection, procedural problems and computational verification and advanced topics. They were moderately reliable for circuit analysis, diagram interpretation but less reliable for foundational concepts, physical intuition or troubleshooting. Grok 4.1 is inconsistent across all categories. It is not recommended for educational use without thorough oversight. The high variability makes it unpredictable.

Instructors are encouraged to use AI as a teaching tool, not a replacement; demonstrate AI strengths and weaknesses in class, show students how to verify AI outputs and discuss why AI fails on certain problems. They should design AI resistant assessments by emphasizing circuit drawing and sketching while solving problems, include conceptual explanation requirements, use in-class problem-solving with physical reasoning and establish clear usage policies.

Students are encouraged to use AI strategically to verify calculations, explore alternative solution methods or to understand conceptual explanations. It should not be used to copy solutions without understanding, skip learning fundamental concepts or trust AI for diagram interpretation without checking. Focus on understanding concepts is encouraged. Students should ask AI “why” and “how” questions, request for explanation not just asking for solution. Instead of asking for solutions, try solving the problems independently first and use AI for verification.

Curriculum designers should restructure assessment weightings by increasing in-class exam weighting, reduce take-home homework weight, include oral/presentation elements and add lab/practical component. Enhanced lab components will facilitate the connection of theoretical knowledge to practical by increasing hands-on circuit building. They should add more open-

ended design problems, require circuit sketching and annotations, and include troubleshooting scenarios.

Future research should expand the domain by including broader engineering subjects, interdisciplinary problems and mathematical foundations. It should explore iterative prompting to see if the AI models can correct themselves after a wrong calculation or track model evolution by re-running these evaluations on latest models.

References

[1] OpenAI, “Introducing GPT-5.” Aug. 07, 2025. [Online]. Available: https://openai.com/index/introducing-gpt-5/

[2] Y. Hao, Y. Liu, B. Liu, G. Amarantidis, and R. Ghannam, “Integrating AI in Engineering Education: A Comprehensive Review and Student-Informed Module Design for U.K. Students,” *IEEE Trans. Educ.*, vol. 68, no. 2, pp. 173–185, 2025, doi: 10.1109/TE.2025.3536105.

[3] L. M. Sánchez-Ruiz, S. Moll-López, A. Nuñez-Pérez, J. A. Moraño-Fernández, and E. Vega-Fleitas, “ChatGPT Challenges Blended Learning Methodologies in Engineering Education: A Case Study in Mathematics,” *Appl. Sci.*, vol. 13, no. 10, 2023, doi: 10.3390/app13106039.

[4] Y. Lukhmanov, A. Perveen, and M. Tsakalerou, “ChatGPT in Engineering Teaching & Learning: Student and Faculty Perspective,” in *2025 IEEE Global Engineering Education Conference (EDUCON)*, 2025, pp. 1–5. doi: 10.1109/EDUCON62633.2025.11016423.

[5] I. Mekterović and Lj. Brkić, “Can LLM Pass a CS1 Course?,” in *2025 MIPRO 48th ICT and Electronics Convention*, 2025, pp. 216–221. doi: 10.1109/MIPRO65660.2025.11131962.

[6] V. Robledo-Rella, A. Gonzalez-Nucamendi, L. Neri, R. M. G. García-Castelán, J. Noguez, and J. Valverde-Rebaza, “Can We Trust AI Chatbots to Teach University Physics? A Performance Comparison of AI Chatbots,” in *2025 IEEE Global Engineering Education Conference (EDUCON)*, 2025, pp. 1–7. doi: 10.1109/EDUCON62633.2025.11016485.

[7] T. Nguyen and H. Sayadi, “ChatGPT vs. Gemini: Comparative Evaluation in Cybersecurity Education with Prompt Engineering Impact,” in *2024 IEEE Frontiers in Education Conference (FIE)*, 2024, pp. 1–9. doi: 10.1109/FIE61694.2024.10893499.

[8] S. Nath and S. Y. Yoon, “WIP: Beyond Code: Evaluating ChatGPT, Gemini, Claude, and Meta AI as AI Tutors in Computer Science and Engineering Education,” in *2024 IEEE Frontiers in Education Conference (FIE)*, 2024, pp. 1–5. doi: 10.1109/FIE61694.2024.10893528.

[9] C. Onuoha, Y. Haga, D. Ferdousi, and T. C. Thang, “Multimodal Large Language Models for High School Mathematical Reasoning: Impact of Input Modality and Artifacts,” in *2025 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, 2025, pp. 1–6. doi: 10.1109/AIMS66189.2025.11229560.

[10] Google Cloud, “Gemini 2.5 Pro.” Dec. 30, 2025. [Online]. Available: https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro

[11] xAI, “Grok 4.1.” Nov. 17, 2025. [Online]. Available: https://x.ai/news/grok-4-1/