Georgia Tech

Georgia Tech Academic AI Playbook

2026 Edition

EDITED BY

David A. Joyner

Nimisha Roy

ISBN 979-8-9978407-0-9
Permanent Link: https://hdl.handle.net/1853/81982

Suggested Attribution
When citing this volume as a whole, please use the following reference (APA 7th edition, the style used throughout the volume):

> Joyner, D. A., & Roy, N. (Eds.). (2026). *Georgia Tech Academic AI Playbook* (2026 ed.). Georgia Institute of Technology. https://hdl.handle.net/1853/81982

When citing or reusing an individual article, please credit that article's author or authors and the volume:

> [Author, A. A.] (2026). [Article title]. In D. A. Joyner & N. Roy (Eds.), *Georgia Tech Academic AI Playbook* (2026 ed., pp. XX–XX). Georgia Institute of Technology. https://hdl.handle.net/1853/81982

Title: Out of Phase: A Declining Rate of AI Error Correction by Power Engineering Students

Authors:

- Samuel Talkington; School of Electrical and Computer Engineering; talks@umich.edu
- Daniel K. Molzahn; School of Electrical and Computer Engineering; molzahn@gatech.edu

Categorization: Teaching AI

Sub-Categorization: Domain-Specific AI Courses and Course Content

---

Formation

ECE 4320: Power System Analysis & Control is the first dedicated course on electric power systems at Georgia Tech, typically taken by junior and senior undergraduates in the ECE Electric Energy Systems Thread and by first-year graduate students. With approximately 30 students enrolled per semester, the course covers the basics of electric power grid analyses including three-phase circuits, the power flow equations, optimal power flow, unit commitment, and state estimation. For many students, this course is both the first and final formal classroom treatment of power systems they will receive before entering the power engineering profession, so what they practice here tends to travel with them.

Between Fall 2023 and today, large language models (LLMs) have progressed from a novelty to an everyday tool in engineering coursework. Their arrival carries particular risk in power engineering: the electric grid is critical infrastructure, and an engineer who accepts a fluent but wrong calculation, a dropped sign convention, a botched phasor operation, or a unit error can carry that mistake into practice at megawatt scale. Early evaluations reinforced this concern; one study found that a majority of GPT-3.5 responses to economic dispatch problems violated basic power balance constraints (Hickey, Ó Faoláin, & Cuffe, 2024).

Rather than trying to ban these tools, we chose to place LLM outputs directly in front of students to assess the students' ability to identify and correct errors. The initiative spanned three course offerings of ECE 4320—Fall 2023, Spring 2025, and Spring 2026—each semester using the most advanced ChatGPT model at the time. We documented outcomes from the first two offerings in a paper published at the *North American Power Symposium,* a conference focused on power systems engineering (Talkington & Molzahn, 2025). Watching the LLM models improve across course offerings along with the corresponding impacts on student experiences turned out to be as instructive as any single assessment.

Action

For each course offering, we wrote an original power factor correction problem from scratch, which ensures that neither the problem nor its solution appears in any model's training data. Power factor correction is a common practical task that requires limited fundamental knowledge and only basic algebraic manipulation, but can pose significant conceptual and engineering design challenges for students. In a representative version, a single-phase voltage source supplies a load whose active power is fixed while its reactive demand varies over a wide range; students select reactances for several capacitors connected to switches along with a rule for the switching configuration so that the worst-case power factor at the source is as close to unity as possible. While formulated to have a single correct solution, these problems task students with balancing tradeoffs in the circuit's performance across a range of parameter values, thus requiring an aspect of engineering design.

We input the same problem to ChatGPT and pasted its full response, unedited, into the first exams of the Fall 2023, Spring 2025, and Spring 2026 semesters. Students were asked to identify every error in the ChatGPT response, explain the physical principle each error violates, and provide a corrected design. The grading rubric awarded separate credit for identification and for correction of errors. Figure 1 shows the Spring 2026 instance of this circuit—a three-element switched capacitor bank—and Figure 2 reproduces the corresponding prompt input to GPT-5. As take-home exams, students had access to any resources at their disposal so long as they worked independently of other students and were given several days to complete the assessments.

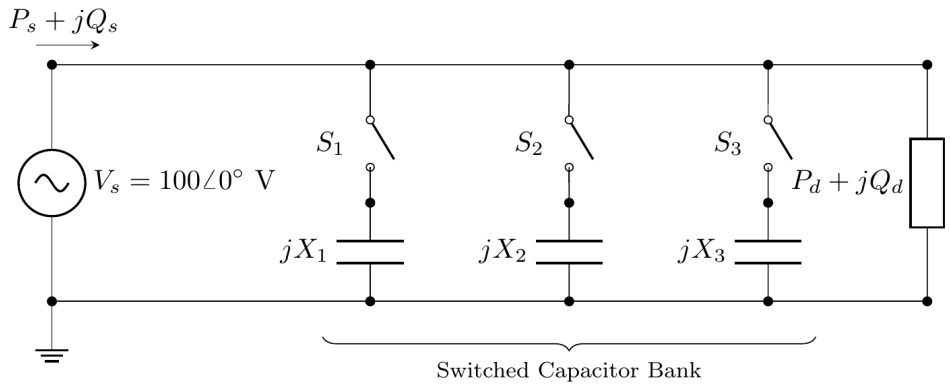

Figure 1: The power factor correction circuit posed to ECE 4320 students and GPT-5 in Spring 2026. A load draws constant active power P_d = 50 W and varying reactive demand $0 \le Q_d \le 60$ VAr, in parallel with a three-element switched capacitor bank (reactances jX_1, jX_2, and jX_3 behind switches S_1, S_2, and S_3).

> Prompt to ChatGPT 5 (Spring 2026)
>
> The following problem for a university-level power systems engineering course explores power factor correction. This problem considers a power system device known as a ‘‘switched capacitor bank’’. This device consists of a set of three capacitors with reactances X_1, X_2, and X_3 that are connected in parallel through three switches S_1, S_2, and S_3 that can be independently switched on or off.
> The switched capacitor bank is connected in parallel with a voltage source that supplies a voltage phasor of $V_s = 100$ V.
> The switched capacitor bank is also connected in parallel with a load demand that consumes a complex power of $P_L + jQ_L$. The active power demand P_L is a constant value of 50 W. The reactive power demand Q_L continuously varies between a minimum value of 0 VAr and a maximum value of 60 VAr.
> Your task in this problem is to determine both the (fixed) values of the reactances X_1, X_2, and X_3 and a switching strategy for the switches S_1, S_2, and S_3. Choose these so that the power factor of the power supplied by the voltage source ($P_S + jQ_S$) is as close to unity (1.0) as possible in terms of its magnitude (leading or lagging) for any value of the load’s reactive power demand Q_L within the range from 0 VAr to 60 VAr. For each switching configuration, give the worst-case power factor of the power supplied by the voltage source (furthest in magnitude from 1.0, either leading or lagging). Indicate whether this worst-case power factor is leading or lagging.
> Note that the reactances correspond to capacitances, so the reactance values X_1, X_2, and X_3 are negative quantities.

Figure 2: The prompt given to GPT-5 for the Spring 2026 power factor correction problem.

The character of the model output changed dramatically across offerings. In Fall 2023, GPT-4 produced roughly a dozen fundamental mistakes: GPT-4 applied an incorrect impedance formula, assigned a positive reactance to a capacitor (sign error), dropped orders of magnitude in its arithmetic, and invented target power factors the problem statement never mentioned. Conversely, in Spring 2025, the o1 model produced a fluent, internally consistent design containing only three subtle flaws. The most important flaw was a design choice that placed the unity power factor operating points at the extremes of the reactive demand range rather than inside it, sacrificing worst-case performance. Students accordingly faced a different task in each offering: catching relatively obvious bookkeeping and arithmetic mistakes in the first versus auditing a plausible but suboptimal engineering argument in the second.

The Spring 2026 offering again included a problem of this format on the first exam, worth 25 points (a quarter of the 100 points in total) split between error identification (10 points) and correction (15 points). No change to syllabus policy was required for this activity, since the model output is itself part of the assessment. However, one design obligation deserves emphasis: newer reasoning models now solve all of our earlier problems correctly (see the arXiv companion to Talkington & Molzahn, 2025), so every course offering requires a fresh problem and newly generated output from the newest available model.

Result

Roughly ninety students completed these assessments across the three semesters of course offerings (Table 1). The clearest change is in the model's own output: the number of distinct errors in the ChatGPT solution fell from eleven in Fall 2023 (GPT-4) to three in Spring 2025 (o1) to a single subtle yet important error in Spring 2026 (GPT-5). Student error identification did not move in one direction. The share of students catching every error was 75%, then 37%, then 63%, tracking how blatant each model's mistakes happened to be rather than any steady trend. Error correction, by contrast, declined monotonically: the share of students producing a fully correct design fell from 36% to 27% to 17% (Figure 4). Consistent with our engineering judgement, this suggests that the problems themselves were more challenging over time as the LLM capabilities improved.

Importantly, we report this as an observation. This is not a controlled statistical result, for a variety of factors. Subtle changes in the problem, model, cohort, and grader all changed each term and each offering had only about thirty students. Since we anticipated before collecting the data that fully correcting increasingly subtle errors would grow harder, a one-sided trend test is appropriate. This trend test is consistent with a decline (Cochran–Armitage $p = 0.047$, two-sided $p = 0.093$), as is a logistic regression in which the odds of a fully correct design fall by roughly 39% per offering. We emphasize that the observed statistically significant decline is specifically in the *share of students who comprehensively corrected AI errors.*

Figure 3 illustrates the Spring 2026 error itself, where GPT-5's binary capacitor weights leave a worst-case reactive power mismatch of 5 VAr compared to 4 VAr for the best possible weights. While the conceptual underpinnings are important for students to master, we note that the implications of this error would be limited in practice.

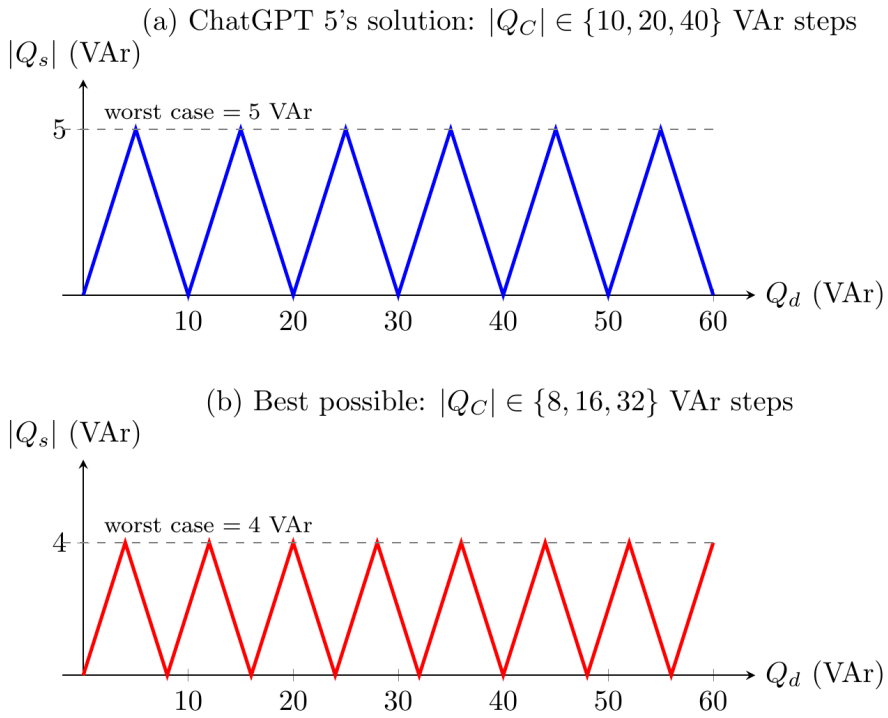

Figure 3: Reactive power mismatch $|Q_s|$ vs. reactive power demand $|Q_d|$ of the load. GPT-5's binary capacitor weights (a) produce a worst-case mismatch of 5 VAr; the best possible weights (b) reduce it to 4 VAr.

Table 1: Student outcomes on the LLM error identification and correction assessments in ECE 4320.

Course Offering	GPT model (# errors)	Full error identification	Full error correction
Fall 2023 (n = 28)	GPT-4 (11)	75% (mean 13.7/15)	36% (mean 7.5/10)
Spring 2025 (n = 30)	o1 (3)	37% (mean 8.7/10)	27% (mean 11.9/15)
Spring 2026 (n = 35)	GPT-5 (1)	63% (mean 9.1/10)	17% (mean 10.6/15)

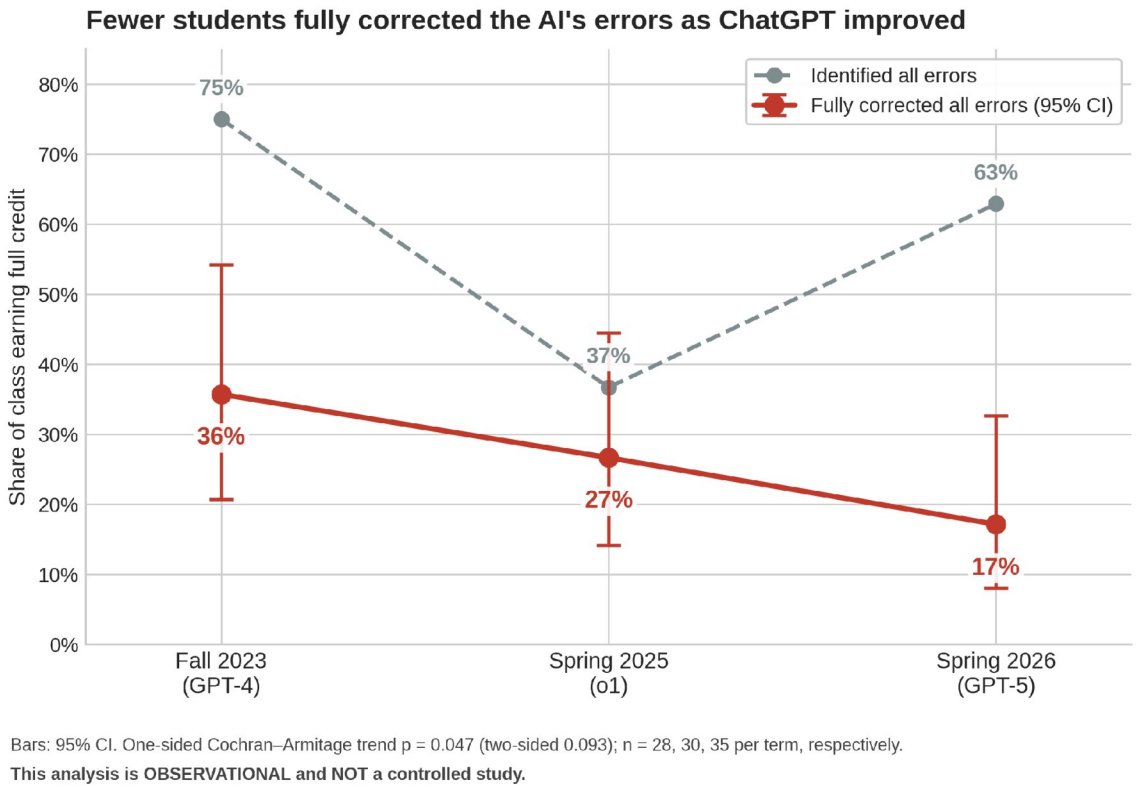

Figure 4: Share of students earning full credit for error identification (grey) and error correction (red) vs. semester of enrollment, with 95% confidence intervals. Full correction declines monotonically (36% → 27% → 17%); identification does not. This analysis is observational and is not a controlled study.

The observation that stays with us most is that the errors of newer models resemble the misconceptions of a capable student. In Spring 2026, several students adopted GPT-5's exact binary-weighting method and then reproduced the same suboptimal choice the model made, a convergence visible directly in the grading rubric. One Spring 2025 student reported independently solving the problem prior to looking at the LLM output, arriving at the same incorrect answer ChatGPT gave, and then not knowing how to proceed. As model output grows more plausible, distinguishing a copied LLM error from a student's own misunderstanding becomes harder, which complicates both grading and academic integrity conversations. While this observation comes from a retrospective account of one course rather than a generalizable measurement, we believe that it may be increasingly relevant across various disciplines.

Lessons

- **Lesson 1: Effectiveness.** We found that tasking students with assessing and correcting an LLM's output is a very effective pedagogy, and we intend to continue this practice. Asking students to audit a confident, polished solution rewards the verification habits practicing engineers need while sidestepping academic integrity and enforcement questions because the LLM output is part of the assessment itself. In this context, writing every problem from scratch is essential: an original problem guarantees the model is reasoning rather than reciting, and it also guarantees students cannot find the answer where the model found it.
- **Lesson 2: Rapid Evolution.** Model capability moves faster than the academic calendar. An exercise tuned to one model's failure profile can be obsolete by the next offering: GPT-4's errors were glaring, o1's were subtle, and current reasoning models frequently correctly solve our

problems outright with a level of detail comparable to the best student solutions. We therefore recommend regenerating the model output with the latest available LLM every term, recalibrating the rubric to the errors that actually appear, and phrasing any published findings as a snapshot of a moving target rather than an enduring claim about what LLMs can or cannot accomplish.

- **Lesson 3: Planning for LLM Improvements.** Based on the rapid improvements in LLM capabilities, we also recommend planning for the day that these models no longer make convenient or obvious mistakes on introductory-level topics. In power engineering, that day appears to be close. The task of creating problems which LLMs fail to solve correctly yet undergraduate students can be reasonably expected to answer becomes harder every semester. This motivates an increasing focus on in-person, collaborative, and formative assessments, where the reasoning happens in view of the instructor. Related lessons include the importance of creating take-home assignments under the assumption that students have full access to LLMs and treating disciplined verification, rather than the spotting of any particular error, as a more meaningful durable skill.

References

Hickey, A., Ó Faoláin, C., & Cuffe, P. (2024). Large language models in power engineering education: A case study on solving optimal dispatch coursework problems. *Proceedings of the 21st International Conference on Information Technology Based Higher Education and Training (ITHET)*.

Talkington, S., & Molzahn, D. K. (2025). VArsity: Can large language models keep power engineering students in phase? *Proceedings of the North American Power Symposium (NAPS)*. https://ieeexplore.ieee.org/document/11272354; An extended and updated version is available at https://arxiv.org/abs/2507.20995

Georgia Tech

ISBN 979-8-9978407-0-9

9 798997 840709

PERMANENT LINK:

https://hdl.handle.net/1853/81982