Is Solving Better Than Evaluating GenAI Solutions?

Ethan Dickey, Marios Mertzanidis, Alexandros Psomas

Purdue University

PURDUE UNIVERSITY

Department of Computer Science

Background and Goal

Generative AI is now ubiquitous in computing courses, but learning outcomes depend on how we integrate it. Instead of asking students to "solve using genAI," we ask: can students learn by evaluating AI-generated algorithm solutions - i.e., acting like a TA who must judge correctness and reasoning?

Research Questions

RQ1. How does evaluating GenAI-generated solutions (vs. solving from scratch) relate to performance on summative assessments (midterm, final)?

RQ2. Does GenAI-evaluation relate to transfer on exam items aligned with prior homework tasks?

RQ3. How do students perceive GenAI-evaluation, and is it associated with self-reported study-habit changes?

Study Design and the Activity

Study design (N = 220)

- Context: Junior-level Analysis of Algorithms (induction, D&C, DP, graphs, flow, reductions).
- Grouping: Students self-reported study pods (≤6); pods randomly assigned to Group A or B to reduce cross-condition contamination.
- Crossover: For HW1-3, Group A did GenAI-evaluation while Group B solved; after the midterm, roles reversed for HW4-6.
- Dosage: Each HW had 4 exercises. Exercises 1-3 were identical across groups (30 pts each). Exercise 4 was the only condition-dependent item (10 pts), deliberately chosen to elicit plausible, instructive GenAI errors.

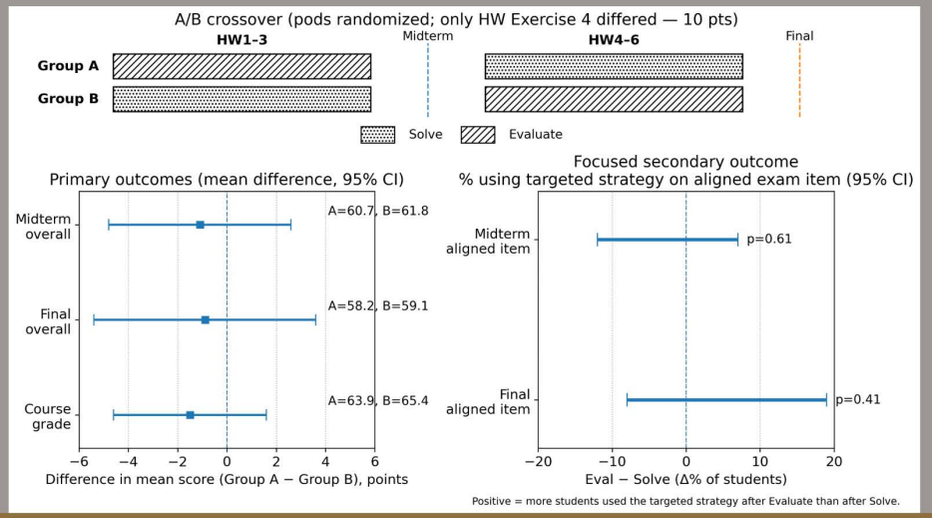
GenAI evaluation task

Students used ChatGPT-4o to generate a succinct solution, then submitted a share-link + screenshot and evaluated the solution:

- Is the algorithm correct? If not, identify the first incorrect step/claim and explain what fails.
- Is the proof/correctness argument valid? If there is a bug, classify it as minor vs. major and justify.

Grading/implementation rigor

Instructor + 7 GTAs + 17 UTAs graded with aligned criteria per exercise (Gradescope). Exercise 4 (the most challenging) was graded by GTAs.



Key Results

No detectable differences on summative performance:

- Midterm (0-101): Group A 60.7 vs Group B 61.8 (95% CI for mean diff includes 0).
- Final (0-113): Group A 58.2 vs Group B 59.1 (95% CI includes 0).
- Course total (%): Group A 63.9 vs Group B 65.4 (95% CI includes 0).

Effect sizes were negligible (|d|< .15), letter-grade distributions did not separate.

Transfer/aligned items:

- Exam items aligned to prior GenAI-evaluation tasks showed no advantage for either condition (midterm had a floor effect; final improved, but not by condition)

Homework patterns:

- Homework differences tracked topic difficulty more than condition (DP/flow harder; earlier topics easier), suggesting syllabus difficulty - not "evaluate vs solve" - drove most homework deltas.

What GenAI got wrong

Across the six topics, GenAI solutions were often plausible but systematically wrong:

- Most often: wrong algorithm design (4/6 topics)
- Sometimes: flawed correctness proofs (2/6 topics)

Student Perceptions

- Helpfulness rating: median 3/5 ("neutral" overall).
- Study habits: 77% no change; 16% "kind of"; 7% yes.
- Students who reported changing study habits rated GenAI-evaluation significantly more helpful (association $p < .001$), suggesting reflective engagement may be a key mediator.

Interpretation

Bottom line: **incorporating structured GenAI-evaluation did not harm achievement in a rigorous setting, but it also did not produce measurable gains on traditional assessments.**

A likely design gap: students often detected errors ("DP state is wrong," "proof skips a base case") but did not repair them - *diagnosis without remediation*.

Design Takeaways for Instructors

If you try this activity (also pick up a handout):

- Start modestly: 1 GenAI-evaluation problem per homework; keep the rest traditional.
- Use a short rubric emphasizing conceptual soundness, algorithm design, and proof validity.
- Make AI outputs contain plausible, instructive errors (missing base cases, misapplied recurrences, flawed reductions).
- Add a required "repair + reflect" step (rewrite the corrected pseudocode; summarize the misconception).

Limitations

- A late change to one homework (subset-sum → subset-product) may have shifted perceived difficulty.
- One final exam problem inadvertently allowed an easier solution path.
- The exposure window was short (12-week intervention); longer and more scaffolded cycles may yield different effects.

Materials & Contact

QR/handout: prompts • rubric • flawed-but-plausible GenAI examples • instructor checklist. • Email: dickeye at purdue.edu