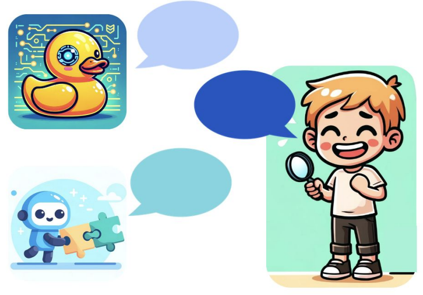


AI Mentors for Student Projects:
Spotting Early Issues in Computer Science Proposals

Gati Aher, Robin Schmucker, Tom Mitchell, Zachary C. Lipton
Carnegie Mellon University

Summary

Problem: Evaluating project proposals and student readiness for project-based learning (PBL) is time-consuming and difficult to scale. Quick and early evaluation could help educators determine which students need support to thrive in PBL

Findings from user study (n=36):

→ GPT-4o's ratings of project proposals show high agreement with educator ratings

→ Novices struggle with writing high-quality proposals and learning objectives

→ Users perceived the system as helpful for writing project proposals and identifying tools and technologies to learn more about

Task & Evaluation Criteria

Project Proposal Task: adapted from existing high school CS Career Pathways PBL program

- **Topic** (Web Dev, Video Game Dev, Useful Scripts)
- **Background / Problem, Objectives**
- **Find + Analyze Inspiring Examples**
- **Evaluation Plan**
- Name 3 skills to develop by working on project
- Pair each skill with mentor (I would like advice from…), job tasks, and technologies (*source: O*Net Online*)

Evaluation: Each project proposals was independently evaluated by two human experts (college CS teaching assistants) and GPT-4o according to a **29-item rubric**

- **1-item x 3 Skill Quality Classification:** Is skill good? (NOT irrelevant to project, non-technical, or vague)
- **3-item x 3 Skill Pairing Classifications:** Is pairing (mentor-skill, job task-skill, technologies-skill) a *good fit*?
- **10-item Quality Checklist:** adapted from college CS mobile dev. project proposal rubric

Item 1. This proposal describes a specific focus and motivation (beyond describing the project topic)	**Item 6.** The design hypothesis describes a specific feature that will be built in the project
Item 2. This proposal describes a good use of computer science skills	**Item 7.** The predicted effects of the design hypothesis can be tested quickly
Item 3. This proposal describes specific tangible features that will be built in the project	**Item 8.** The project has an objective measure of success or learning
Item 4. Working on the project is relevant to learning about project topic	**Item 9.** The design hypothesis has an objective measure of success or learning
Item 5. The proposal analyzes similar products, papers, or applications	**Item 10.** The evaluation plan can be carried out within a 4 week sprint

- **1-item Recommend for Resume:** General measure for quality of project (goes beyond tutorial, understandable)

Educators vs. GPT-4o Evaluate Project Proposals

Skill Quality: High Agreement

Skill Pairings: Low Agreement (hard to grade vague skills)

Quality Checklist: Slightly Low Agreement (GPT-4o is stricter)

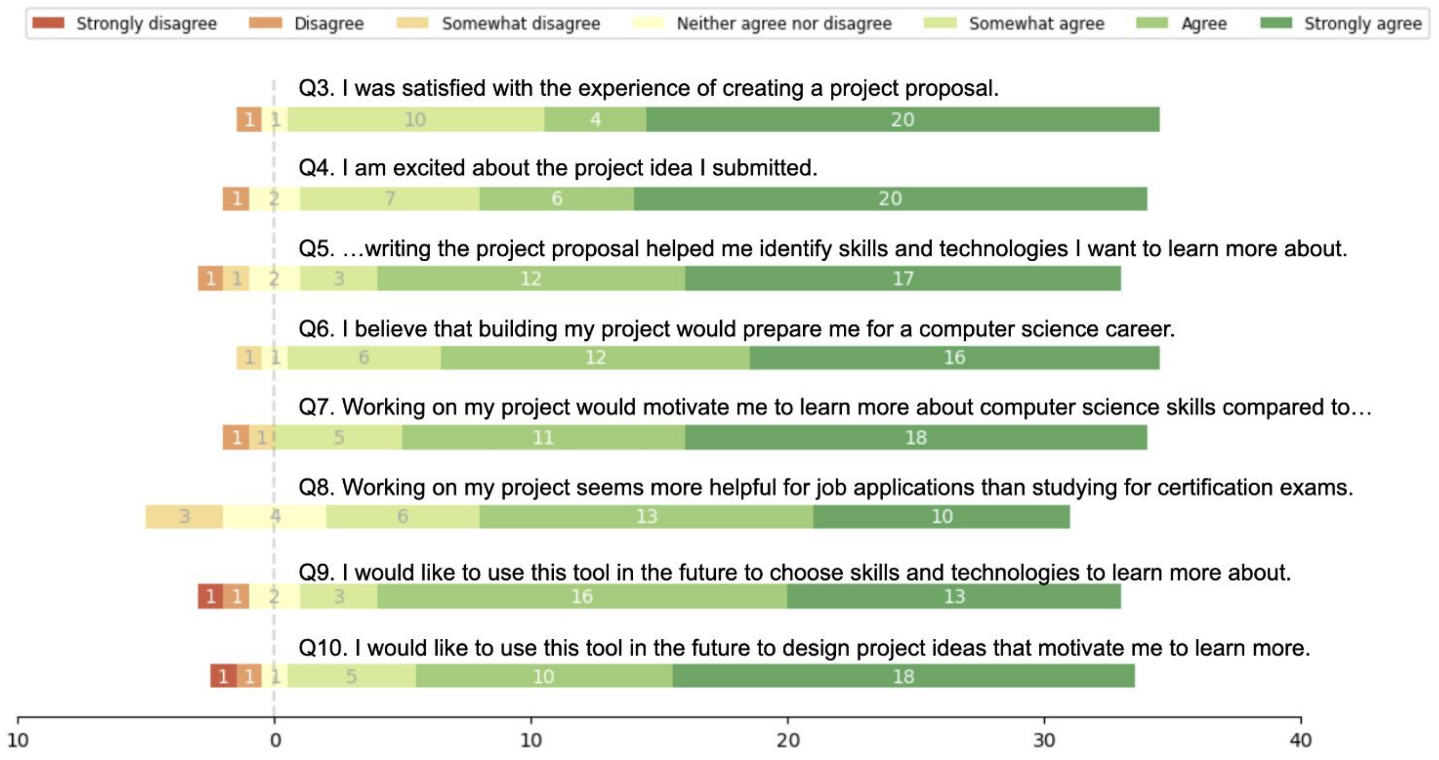
	TA1 / TA2	TA1 / GPT-4o	TA2 / GPT-4o	Avg. κ
Skill Quality Classification	86.1%, $\kappa = 0.72$	84.3%, $\kappa = 0.68$	81.5%, $\kappa = 0.63$	0.68
Skill Pairing Classification	68.6%, $\kappa = 0.26$	74.2%, $\kappa = 0.46$	67.2%, $\kappa = 0.20$	0.29
Quality Checklist	87.2%, $\kappa = 0.49$	71.4%, $\kappa = 0.28$	74.7%, $\kappa = 0.28$	0.38
Recommend for Resume	80.6%, $\kappa = 0.50$	75.0%, $\kappa = 0.43$	77.8%, $\kappa = 0.45$	0.46

	TA 1	TA 2	GPT-4o	Self
TA 1	1.00	0.74	0.70	0.16
TA 2	0.74	1.00	0.53	0.38
GPT-4o	0.70	0.53	1.00	0.23
Self	0.16	0.38	0.23	1.00

Spearman correlations show that though GPT-4o is a *stricter* grader than the TAs, **it's grades preserves the rank order of the Human Experts' Quality Checklist scores** (better than students' self-evaluations)

Task	Experience	TA1	TA2	GPT-4o	Self-Rating
Skill Classification	Novice	35.3%	39.2%	31.4%	-
	Experienced	70.2%	68.4%	68.4%	-
Skill Pairing Classification	Novice	60.8%	86.9%	57.5%	-
	Experienced	53.7%	92.4%	64.9%	-
Quality Checklist	Novice	83.5%	82.4%	58.2%	85.9%
	Experienced	84.7%	90.0%	71.1%	95.8%
Recommend for Resume	Novice	58.8%	70.6%	58.8%	-
	Experienced	78.9%	84.2%	73.7%	-

Sanity Check — Students with prior CS knowledge tend to have higher quality submissions, e.g., on average, novices can propose 1 good skill, experienced students can propose 2 good skills.



Majority of students enjoyed completing the project proposal activity and believed it benefited their learning.

Future Directions

GPT-4o achieves high agreement with human expert evaluations and hence enables scalable support for PBL

→ **Classroom studies:** Does LLM agreement transfer to teachers' rubrics? Student-led rubrics?

→ **Long-term studies:** Does early detection of issues spotted by GPT-4o lead to a smoother, more educational PBL experience?