Examining Teacher-AI Teaming Through the Use of a Generative AI Assessment Creation Tool in High School Mathematics Classrooms

Shuchi Grover[a], Corrin Clarkson[b], Rick Lynch[a], & Brittany Runge[b]
[a] Looking Glass Ventures (LGV) | [b] Indiana University (Mathematics Department)

Objectives

As AI and specifically large language models (LLMs) and generative AI (GenAI) tools like ChatGPT (OpenAI, 2023) enter K-12 education spaces, there is a push to center educators in the AI in education revolution (US DOE, 2023). Research is needed to determine ways to supporting collaboration between teachers and AI systems to promote effective and productive uses of AI (NASEM, 2022; OSTP, 2023). This paper presents research that contributes to the understanding of successful teacher-GenAI teaming and what factors impact AI as a change agent in teaching. An additional goal is to understand the new skill of instructing an LLM about one's goals or "prompt engineering" (White et al., 2023).

This project in response to NSF 23-097 is a collaborative effort between the Dept. of Mathematics at Indiana University, Bloomington (IU) and Looking Glass Ventures, LLC (LGV), the developer of 'Edfinity' (an NSF-supported homework system and assessment platform). The project seeks to research the use of Edfinity's new LLM module, ALICE (A Language-Independent Codeless Environment; Edfinity, 2023), in IU's Advance College Project (ACP). The ACP is a dual enrollment partnership accredited by the National Alliance of Concurrent Enrollment Partnerships between IU and selected high schools throughout Indiana and surrounding states. Edfinity is used in formative assessment for IU's math courses and is slated to be used by those in ACP. The goals of this project are to (1) augment teacher capabilities, (2) empower teachers to author technology-enhanced assessments (TEAs), thereby ensuring the assessments meet their students' needs, (3) examine and foster AI literacy and trust among teachers, (4) alleviate teacher workload to accommodate more students and with differentiation, (5) contribute to the understanding of successful teacher-AI teaming through research conducted by an interdisciplinary team involving high school math teachers, math professors, learning sciences and HCI researchers, and AI specialists and (6) empirically examine and develop the ***science of domain-specific prompt engineering***.

Our efforts align with Strategy 2 of The National Artificial Intelligence R&D Strategic Plan (NAIRD; OSTP, 2023), focusing on effective human-AI collaboration. We address various strands in the NAIRD, including developing the science of human-AI teaming, articulating a model of team performance, cultivating trust in human-AI interaction, understanding human-AI interactions, and fostering teacher collaboration with AI. TEAM AI also demonstrates responsible AI R&D collaboration between academia and industry (Strategy 8), exploring the practical implementation of AI advances in a high school math classroom.

In this paper we describe our research on the integration of ALICE into high school Finite Mathematics and Calculus courses. Edfinity assessments utilize the popular, open-source WeBWorK format (Gage, Pizer, & Roth, 2002) to deliver interactive, auto-gradable, isomorphic TEAs to support classroom assessment for better student learning. ALICE works in conjunction with the OpenAI API, utilizing GPT-4 (OpenAI, 2023) trained on a large corpus of existing assessments to generate WeBWorK source code (in the PERL programming language). Given natural language prompts from teachers, ALICE converts the natural language specifications

into structured queries for the AI model. These queries are then used to generate math problems along with the corresponding WeBWorK source code for an interactive, isomorphic assessment along with hints and a solution (Figure 1). Such code would otherwise have to be written by programmers and effectively left K-12 teachers out of the equation of creating WeBWorK assessments for themselves.

Background and Research Framework

Teachers are increasingly utilizing generative AI chatbots like ChatGPT to create homework prompts, adapt instructional content, generate lesson plans, and assist with administrative tasks (Marr, 2024), as well as automate grading, provide personalized feedback, and offer real-time practice in subjects like mathematics and foreign languages (Dorn et al., 2023). Despite this growing adoption, opinions on AI tools in education are mixed.

Our teacher-AI teaming designs leverage Pea's ideas of intelligence distributed (Pea, 1993) "across minds, persons, and the symbolic physical environments, both natural and artificial" (p. 47). We aim to study the re-configuration of this AI- augmented socio-technical system where TEAs that used to be coded by a programmer and provided to a teacher to use with few opportunities of communication between them, are now created through teachers determining what problems should be created and the AI actualizes the creation of the TEA.

Our research is focused specifically on classroom formative assessment with GenAI. Teachers' use of formative assessment provides them insight into students' understanding, which in turn, helps them identify student misconceptions (Heritage and Wylie, 2018). Recent attention to equity and deeper learning (Pellegrino & Hilton, 2013) has prompted a transformation in technology-based formative assessments (Conley & Darling-Hammond, 2013). The use of rich, readily scorable math assessments also makes possible timely formative feedback which is beneficial, especially for struggling students (Babaali & Gonzalez, 2015; Sehran, 2019). Immediate feedback has been found to be more effective than delayed feedback (Van der Kleij et al., 2015), especially for low-ability learners or difficult tasks (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Shute 2008) which makes TEAs on platforms like Edfinity better for equitable outcomes in math classrooms. Additionally, in the context of math teaching and learning, isomorphic or "structurally related mathematical problems" (the kind that ALICE will help teachers create without writing any code) have been shown to help students develop a conceptual understanding (Greer and Harel, 1998) and to uncover student misconceptions and error patterns (Attali & van der Kleij, 2017; Kusairi, Alfad & Zulaikah, 2017).

The following questions guided this research:

1. What are teachers' experiences as they partner with ALICE to generate TEAs? What factors shape these experiences?
2. What are teachers' experiences with prompt engineering in ALICE to support their formative assessment needs in high school Finite Mathematics and Calculus 1 courses?
3. What are these teachers' views on AI? How are they impacted by this experience?

Methods and Data Sources

Teachers from rural, urban, and suburban high schools across Indiana teaching in the dual-enrollment program of a large research university were recruited at the start of AY 2023-2024. They were provided a link to a pre-survey (along with an informed consent form to opt in to the research) with Likert-scale and open-ended questions probing teachers' understanding of and views on AI in education (adapted from Cukurova et al., 2023). Teachers who completed the pre-survey (17 in the Fall and an additional 7 in the Spring) were invited to

attend a 1-hour training session facilitated by the researchers, in which they were provided some background on how LLMs work, how ALICE was trained, how to write prompts for ALICE, and research participation tasks. Teachers were requested to submit 1-2 problem prompts per week to generate problems and capture their feedback on the problems generated and (optionally) their use with students in an individual teacher log. Teachers were provided access to a shared corpus (on Edfinity.com) of problems generated by all the teachers along with the prompts. The post-survey at the end of each term included the same questions on AI attitudes to capture change from pre-to-post, as well as Likert-scale and open-ended questions pertaining to the ALICE experience.

Our data sources include:

- A corpus of over 400 teachers' prompts and math problems generated for topics in Finite Mathematics, Brief Survey of Calculus I, and Calculus 1.
- Teacher logs with reflections on their satisfaction with the prompt, use with students, and (optional) re-submission with tweaks.
- Pre-post teacher surveys.
- Semi-structured interviews with about half the teachers in each term (6 in the Fall and 7 in the Spring). Teachers were invited to opt-in for the interview.

Eleven teachers in the Fall and three in the Spring completed data items 1-3 above for a total N of 14. **All 11 teachers from the Fall participated in the Spring study** (Table 1).

In an interesting research methodology experiment, the researchers also used the LLM ChatGPT4o as a "research partner" to help code the interviews and some open-ended survey responses (*as described in a separate paper (Grover, in review)*).

Results

Findings from Fall 2023

Teachers' experiences with ALICE were largely positive. A qualitative analysis of their feedback to the answers to the question "How would you describe your experience to other teachers?" involved coding (by a researcher and an LLM with over 85% inter-rater reliability). The following themes emerged from the analysis:

- **Enhanced Efficiency:** Teachers liked having a tool to create numerous similar (isomorphic) problems effortlessly.
- **Ease of Implementation:** The problems received back from ALICE were easy to use in the classroom and could be quickly integrated into lessons and with students.
- **Empowerment for Teachers:** ALICE was viewed as a valuable tool in the teacher's toolkit, contributing to improved learning outcomes. It offered freedom to create problem sets aligned with standards and concepts.
- **Challenges and Learning Curve:** Teachers encountered limitations in the variety of questions produced for specific topics. While ALICE's generated questions did not always meet the desired level of rigor, the variety and accuracy are appreciated.
- **Attitude Toward AI:** Some teachers initially had reservations about AI but found ALICE to be user-friendly and accessible.

In semi-structured interviews with 5 teachers, it emerged that a key area of concern and interest was "prompt engineering"— how they were interacting with ALICE to get desired results. We encouraged teachers to experiment with refinements of prompts in the Spring 2024 term.

Findings from Spring 2024

Teachers' largely positive experiences continued from the Fall to the Spring semester. In the post-survey, 12 of 14 teachers responded "4" (agree) or "5" (completely agree) with the remark "I *have enjoyed this experience of using ALICE to generate questions for use in my course*." 11 of 14 responded with 4 or 5 to "*Working with ALICE helped me with generating creative word problems*." 13 of 14 teachers responded with 4 or 5 to "*This experience has made me aware of uses of AI that I had not thought of before*." 13 of 14 teachers responded with 4 or 5 to "*Uses of generative AI like ALICE have the potential to positively impact classroom teaching."* Table 2-4 share teachers' responses to Likert-scale and open-ended questions about their ALICE, and specifically prompt engineering, experiences. Table 5 shares how their views on AI changed due to their experience with ALICE.

Analysis of Teacher Interviews

We conducted semi-structured interviews (20-40 minutes long) with seven teachers at the end of the Spring semester. Five of the teachers participated in both semesters, and among them, three had been interviewed at the end of the Fall as well. Five of the seven teachers interviewed had very positive experiences.

A qualitative analysis of the interviews (based on frequency and depth/strength of remarks) revealed the following factors as having influenced teachers' overall experience, and Figure 3 shows how important each of these factors was for each teacher interviewed.

– Usefulness in Teaching,
– Prompt Creation and Refinement,
– Student Use & Reactions,
– AI as a Thought Partner,
– Comparison with Usual Approach (to assessments),
– Attitude toward AI and Future Use.

In order to give a richer pictures of the experience, Table 1 presents brief case studies of four of the seven teachers (F5, F6, M1, and M7) selected based on a maximum variation strategy (Flyvbjerg, 2006) to represent diversity in gender and school context, courses taught, and views on the ALICE experience.

Teachers' Understanding and Views on AI in Education

The pre-post survey asked teachers what they understood by the term "artificial intelligence" (Table 7)- a term we had not defined as part of this intervention. An open-coding analysis of the responses revealed more nuanced themes from pre- to post-intervention. The most dominant pre-intervention theme was "automation and the replication of human tasks." Post-intervention, the most dominant theme was "learning and adaptation." Figures 4-6 pre-to-post responses to questions related to teachers' views on AI generally. The responses were broadly same-to-slightly more positive from pre-to-post. The low N precluded a t-test of significance.

Discussion and Scholarly Implications

This project examines the experiences in the use of ALICE by high school math teachers with the goal of contributing to evidence-based teacher-AI teaming.

Although the 14 participating teachers' experiences varied, the results of this implementation study are promising. Survey and interview responses suggest that it was not only a positive experience for 85% of the teachers but also an enlightening one for all the teachers. It opened their eyes to the possibilities of teachers' use of AI. They found value in being able to generate practice problems, especially on topics on which they felt students needed help. Having isomorphic TEAs meant they could produce endless versions of practice problems

for their students, especially for review before summative assessment tests. Interesting, this experience with ALICE did not require them to change how they normally taught. It aided their teaching and alleviated some of the assessment creation- related burden.

Prompt engineering presented an interesting challenge and learning curve for teachers. By and large, they did not find the process of writing prompts to get the desired question to be challenging/difficult. However, getting the desired result from ALICE dominated their experience. As Subramonyam et al. (2024) aver, “While LLMs are capable of interpreting a vast range of queries, their very flexibility can pose challenges for users attempting to convey precise intentions.” They describe the “gulf of envisioning,” which captures the challenge users face in successfully formulating their intentions to elicit the desired response from an LLM. Given that teachers did not find the training to be wanting, and had a very clear idea of what they wanted as the output, it appears that the gap was related more to capability (how to set my goals and intentions such that the LLM can accomplish the task) rather than instruction or intentionality. Teachers’ reflections on the strategies that helped them in prompt engineering suggests the kind of training new teachers could benefit from as they embark on similar uses of GenerativeAI tools, especially in the context of STEM disciplines. Many agreed that they found that "the quality of the generated assessments improved" as they became more experienced with writing prompts.

The nuances of teachers’ reflections in the interviews suggest that responses to questions of whether they will use AI or not in the abstract can be misleading; it all depends on the context. That said, teachers’ prior views on technology and AI do impact their experience and openness (or lack thereof) to try out these new tools (as seen for teacher F6, who had by far the least positive experience). In the future, it would help to understand individual contexts and needs better to find the best ways in which such a tool can help. Interim check-ins with the teachers during the term would have also been helpful to address some of the challenges and issues that emerged in the post-interviews.

The concept of teacher-AI teaming has gained traction as AI technologies, especially LLMs, have become more sophisticated and accessible to teachers. This experience report examines the teacher-Generative AI partnership in high school math teaching, with special attention to teachers’ experiences with using natural language prompts to generate technology-enhanced formative assessment that addresses the learning needs in classrooms. Our results are very promising around the potential of augmenting teachers’ capacity to perform tasks that are time- consuming or simply outside their skill set (such as programming technology-rich assessments). Through attention to the factors that influence teachers’ attitudes toward and use of AI tools and the needs pertaining to the learning curve for writing prompts that generate effective assessments, this implementation study advances our understanding of teacher-AI teaming and prompt engineering. Our findings have implications for the future and on ways to help teachers collaborate effectively with AI tools to enhance their teaching capabilities and improve student learning outcomes. The impact of this early work to examine domain-specific LLMs for teacher use and related teacher preparation will extend beyond Math, as will the positive narrative of centering teachers and extending their capabilities in hitherto inaccessible areas.

Acknowledgements

This research was supported by a grant from the National Science Foundation (#**2335834 and 2335835**).

References

Attali, Y., & van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. Computers & Education, 110, 154-169.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. Review of Educational Research, 61(2), 213-238.

Cukurova, M., Miao, X., & Brooker, R. (2023, June). Adoption of artificial intelligence in schools: Unveiling factors influencing teachers' engagement. In International Conference on Artificial Intelligence in Education (pp. 151-163). Cham: Springer Nature Switzerland.

Diliberti, M., Schwartz, H. L., Doan, S., Shapiro, A. K., Rainey, L., & Lake, R. J. (2024). Using Artificial Intelligence Tools in K-12 Classrooms. RAND.

Edfinity. 2023. https://edfinity.zendesk.com/hc/en-us/articles/15503346805389--video-ALICE-no-code-authoring-of-randomized-algorithmic-problems

Flyvbjerg, B. (2006). Five misunderstandings about case-study research. Qualitative Inquiry, 12(2), 219-245.

Gage, M., Pizer, A., & Roth, V. (2002). WeBWorK: Generating, delivering, and checking math homework via the Internet. ICTM2 International Congress for Teaching of Mathematics at the Undergraduate Level, Hersonissos, Crete, Greece http://www.math.uoc.gr/~ictm2/Proceedings/pap189.pdf.

Greer, B., & Harel, G. (1998). The role of isomorphisms in mathematical cognition. The Journal of Mathematical Behavior, 17(1), 5-24.

Klopfer, E., Reich, J., Abelson, H., & Breazeal, C. (2024). Generative AI and K-12 Education: An MIT Perspective.

Kusairi, S., Hidayat, A., & Hidayat, N. (2017). Web-based diagnostic test: Introducing isomorphic items to assess students' misconceptions and error patterns. Chemistry: Bulgarian Journal of Science Education, 26(4), 526-539.

Marr, B. (2024). Navigating AI for K-12 Educational Efficiency & Effectiveness. Retrieved from Frontline Education. https://www.frontlineeducation.com/empowering-k-12-districts-navigating-ai-adoption-for-enhanced-educational-efficiency-and-effectiveness.

National Academies of Sciences, Engineering, and Medicine (NASEM). (2022). Human-AI teaming: State-of-the-art and research needs. Washington, DC: The National Academies Press. https://doi.org/10.17226/26355.

OpenAI. (2023). *ChatGPT* (May 24 version) [Large language model]. https://chat.openai.com/chat/

Open AI. (2023). GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

OSTP. (2023). National Artificial Intelligence Research And Development Strategic Plan 2023 Update. A report by the Select Committee on Artificial Intelligence of the National Science and Technology Council. Executive Office of the President: Washington, DC, USA.

Subramonyam, H., Pea, R., Pondoc, C. L., Agrawala, M., & Seifert, C. (2024). Bridging the gulf of envisioning: Cognitive challenges in prompt-based interactions with LLMs. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642754.

U.S. Department of Education (US DOE) Office of Educational Technology. (2023). Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations. Washington, DC.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. Review of Educational Research, 85(4), 475-511.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv:2302.11382 [cs.SE].

Figures & Tables

(a) Prompt: Write a question asking to find a derivative of a binomial quadratic function using the definition of derivative where the question shows the solution but requires filling in 1 box per line throughout the process of solving.

Consider the function $f(x) = 5x^2$. Use the definition of the derivative to find $f'(-2)$:

$$f'(-2) = \lim_{h \to 0} \frac{f(-2+h) - f(-2)}{h}$$

a. First, compute and simplify the numerator.

$f(-2+h) - f(-2) =$ ☐

b. Next, factor out h from your answer in Part a.

$f(-2+h) - f(-2) = h($ ☐ $)$

c. Canceling h in the difference quotient gives:

$f'(-2) = \lim_{h \to 0}$ ☐

d. Finally, computing the limit, gives the derivative value:

$f'(-2) =$ ☐

Algorithmic (WeBWorK) Part

```
$a = non_zero_random(-5, 5, 1);
$b = (random(0,1)) ? 0 : random(-5, 5, 1);
$c = ($b == 0) ? random(-5, 5, 1) : 0;
$x_value = non_zero_random(-3, 3, 1);

# Create the function and calculate derivative and value at point
$function = Formula("($a)*x^2 + ($b)*x + $c")->reduce;
$function_xh = $function->substitute(x=>"$x_value + h");
$num_coef = 2*$a*$x_value + $b;
$ans = Compute("$num_coef");

Context("LimitedPolynomial-Strict")->variables->are(x=>"Real", h=>"Real");

$num = Formula("$num_coef h + $a h^2")->reduce;
$factor = Formula("$num_coef + $a h")->reduce;

##########################################################################
#  Problem Statement
##########################################################################

BEGIN_PGML

Consider the function [``f(x) = [$function]``]. Use the definition of the derivat
[``f'([$x_value])``]:
```

(a) **Prompt**: Write a question asking students to graph a rational function in the form (a+bx)/(c+dx), where a, b, c, and d should be whole numbers between -10 and 10. Give students an applet to draw the graph that allows them to select the number of asymptotes, and then graph the function by dragging the asymptote lines and a point on the function.

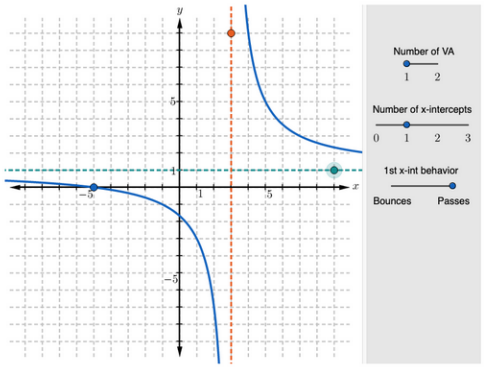
(b)

Plot the rational function: $f(x) = \frac{-7x+7}{4x+9}$

Set the asymptotes first. Both a zero vertical asymptote and zero horizontal asymptote is **not** allowed. All points move in increments of 1/4.

If the horizontal asymptote is nonzero, then there must be at least one x-intercept. Otherwise, there are no x-intercepts and you must adjust the y-intercept.

Depending on the horizontal asymptote being zero or not, you can set whether x-intercept(s) pass through the axis or bounce, or set the multiplicity of the vertical asymptote(s).



Show hint

To plot a rational function, find the x and y intercepts along with the horizontal and vertical asymptotes. Use the multiplicities to determine the behavior of each.

(c)

Algorithmic (WeBWorK) Browse templates

```
Context("Numeric");
Context()->noreduce('(-x)-y');
Context()->noreduce('(-x)-y','(-x)+y');

do {
    do {
        $a = random(2,9)*random(-1,1,2);
        $b = random(2,9)*random(-1,1,2);
        $xint = $b/$a;
    } until (4*$xint == floor(4*$xint));

    do {
        do {$c = random(2,9)*random(-1,1,2)} until (abs($c) != abs($a));
        $d = random(2,9)*random(-1,1,2);
        $va = $d/$c;
    } until (4*$va == floor(4*$va));

    $ha = $a/$c;
} until (4*$ha == floor(4*$ha) && (abs($b) != abs($a) || abs($d) != abs($c)));
$dtop = Formula("$a x - $b")->reduce;
$dbot = Formula("$c x - $d")->reduce;


Context("ArbitraryString");
```

Figure 1. How ALICE works. a) A natural language prompt; b) TEA generated (with a hint) (c) WeBWorK backend code (in PERL programming language) for the rich, isomorphic, interactive, problem type.

Figure 2. Ranked Importance of 6 factors contributing to teachers' experience with ALICE

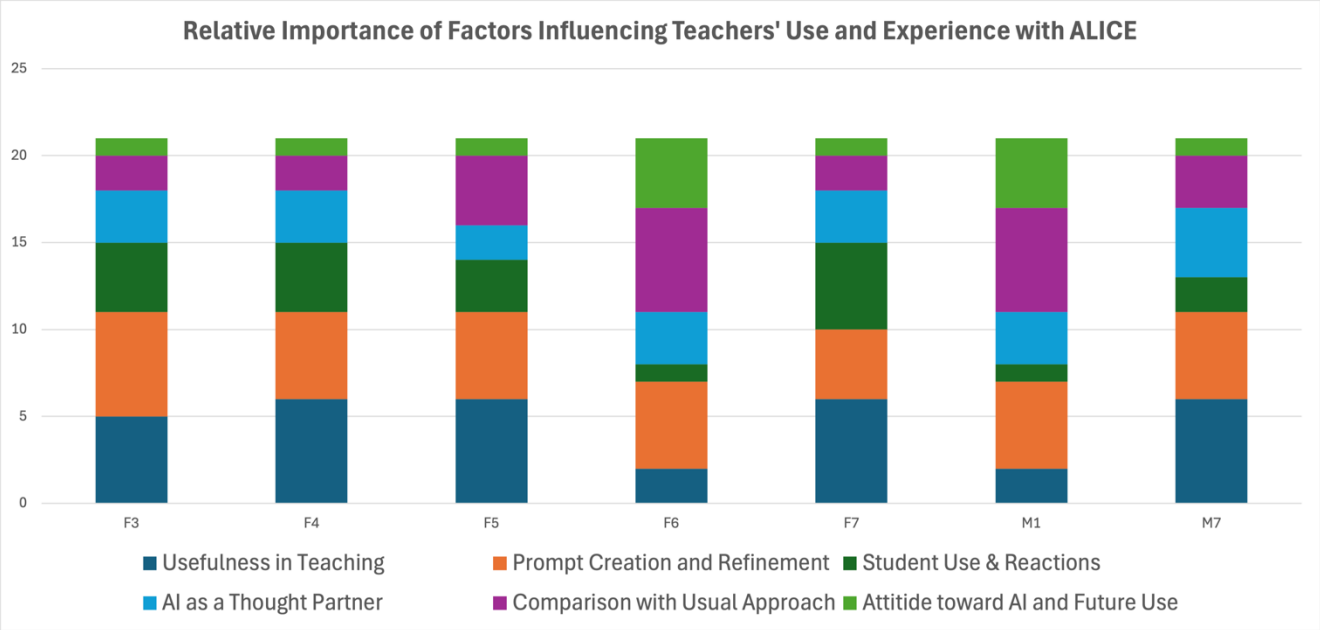


Figure 3. Pre-to-post Likert scale responses to teachers' views on usefulness of AI in education

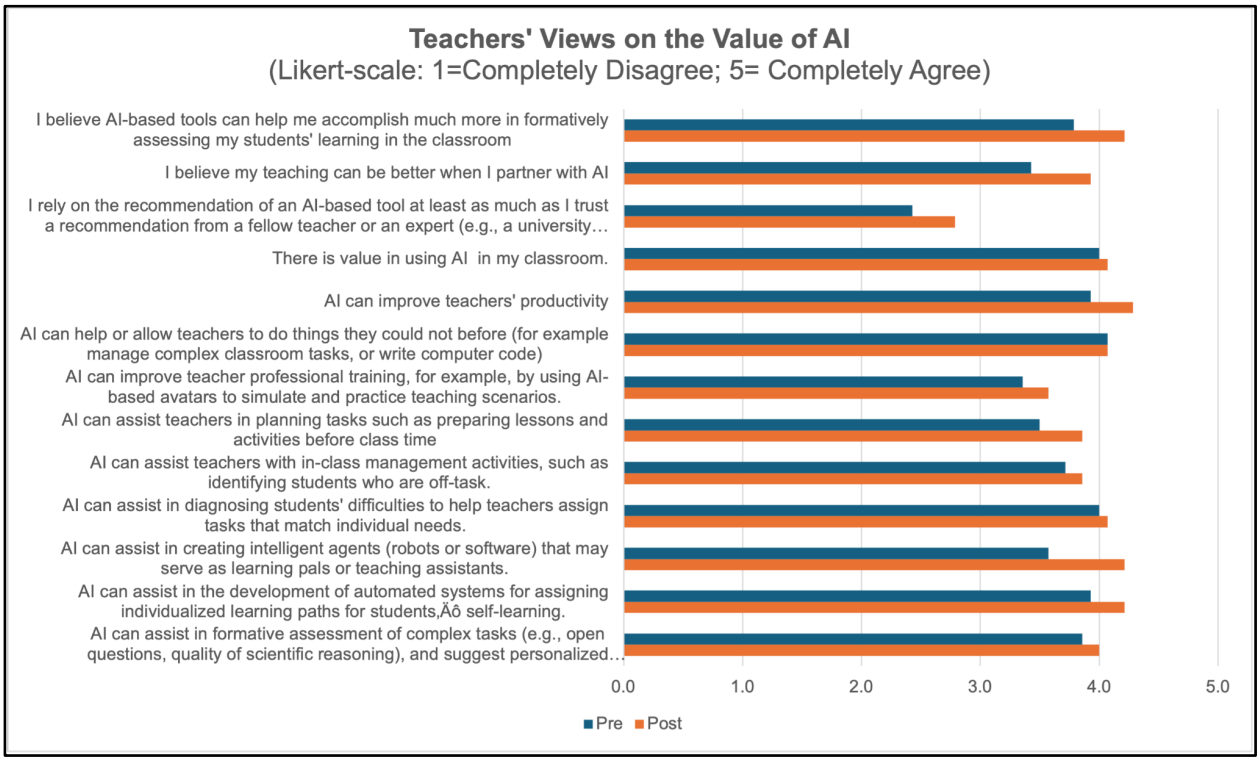


Figure 4. Pre-to-post Likert scale responses to factors that may hinder teachers' use of AI in education

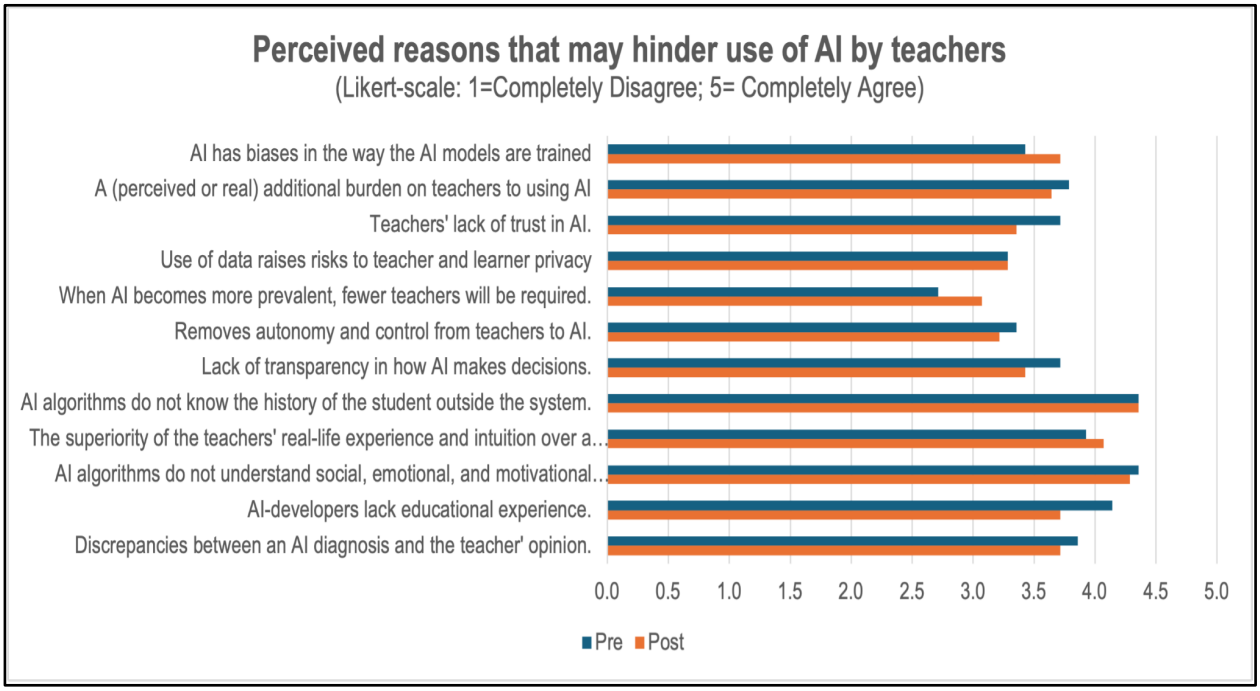


Figure 5. Pre-to-post Likert scale responses to teachers' trust in AI in education

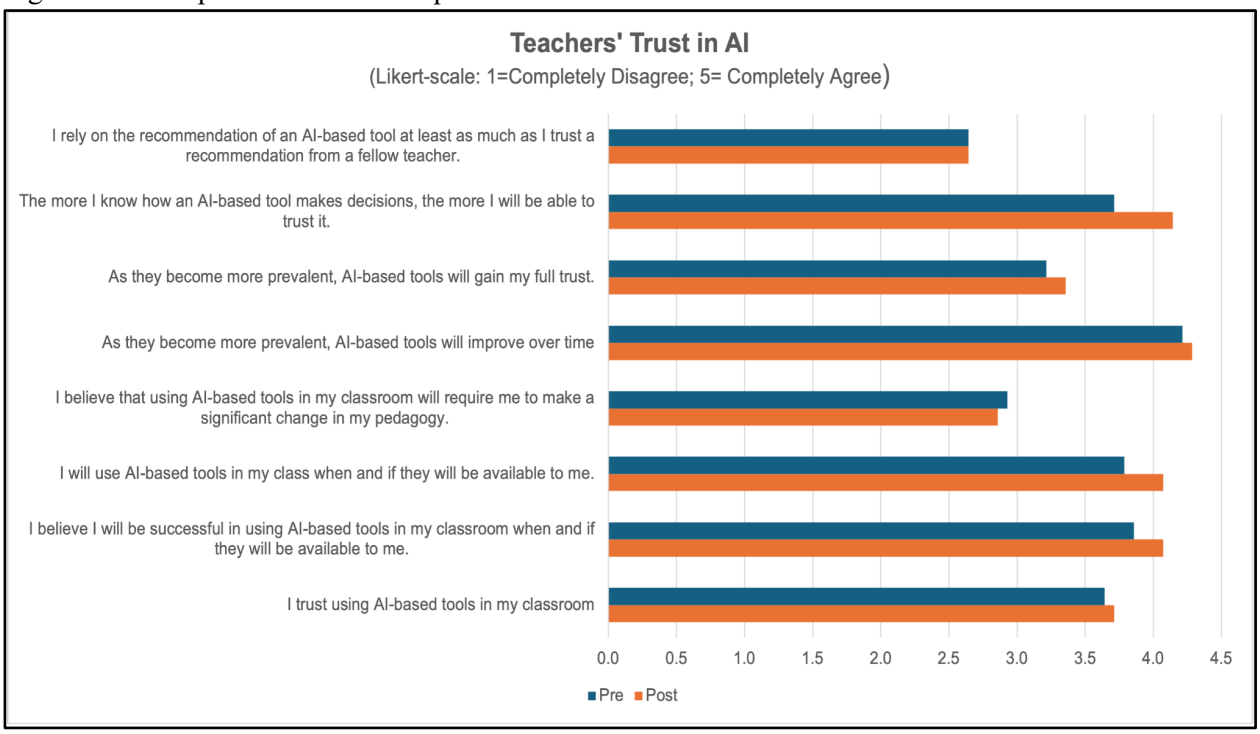


Table 1. Description of the teacher sample (N=14). Shaded cells represent teachers who participated in end-of-semester interview(s). Cells shaded in green are described in the "case studies" in Table 6.

Teacher Code & Gender (Male) **ACP Course(s) Taught** **Research Participation**	**School location** **School size** **Demographics**	**Teacher Code & Gender (Female)** **ACP Course(s) Taught** **Research Participation**	**School location** **School size** **Demographics**
M1 Finite Math Fall (& Interview) Spring (& Interview)	Suburban Indiana Total Enrollment: 2326 White: 44.2%; Black: 9.2%; Asian: 19.6%; Hispanic: 21.8%; Multi-racial: 5% Econ.Disadvantaged: 69.2% Students with Disabilities: 14.1%	F1 Calculus 1 Fall (& Interview) Spring	Rural Indiana Total Enrollment: 1034 White: 94.1%; Black: 0.7%; Asian: 0.5%; Hispanic: 1.7%; Multi-racial: 2.8% Econ. Disadvantaged: 50.8% Students with Disabilities: 15.4%
M2 Finite Math Fall Spring	Suburban Indiana Total Enrollment: 1565 White: 82.2%; Black: 2.6%; Asian: 1.5%; Hispanic: 9.3%; Multi-racial: 4.2% Econ. Disadvantaged: 32.1% Students with Disabilities: 11.9%	F2 Finite Math Fall Spring	Rural Indiana Total Enrollment: 1582 White: 86.2%; Black: 1.8; Asian: 1.4% Hispanic: 6.1%; Multi-racial: 4.2% Econ. Disadvantaged: 32.9% Students with Disabilities: 15.7%
M3 Finite Math Fall (& Interview) Spring	Suburban Indiana Total Enrollment: 2624 White: 84.0; Black: 3.1%; Asian: 4.6% Hispanic: 5.1%; Multi-racial: 2.9% Econ. Disadvantaged: 16.3% Students with Disabilities: 13.9%	F3 Finite Math Fall Spring (& Interview)	Suburban Indiana Total Enrollment: 765 White: 91.2%; Black: 0.7%; Asian: 0.4%; Hispanic: 3.7%; Multi-racial: 3.4% Econ. Disadvantaged: 29.4% Students with Disabilities: 14.5%
M4 Brief Survey of Calc 1 Fall (& Interview)	Suburban Indiana Total Enrollment: 786 White: 82.1%; Black: 5%; Asian: 2.8%; Hispanic: 4.3%;	F4 Brief Survey of Calc 1 Fall	Suburban Indiana Total Enrollment: 3877 White: 71.6%; Black: 10.4%; Asian: 5.6%; Hispanic: 7.3%;

Spring	Multi-racial: 5.9% Econ. Disadvantaged: 29.8% Students with Disabilities: 11.1%	Spring (& Interview)	Multi-racial: 4.4% Econ. Disadvantaged: 18.5% Students with Disabilities: 10.7%
M5 Calculus 1 Fall Spring	Rural Indiana Total Enrollment: 621 White: 88.9%; Black: 1.6%; Asian: 0.2%; Hispanic: 6%; Multi-racial: 3.4% Econ. Disadvantaged: 47.7% Students with Disabilities: 20.8%	F5 Calculus 1 Fall (& Interview) Spring (& Interview)	Rural Indiana Total Enrollment: 1472 White: 93.9%; Black: 0.3%; Asian: 0.7%; Hispanic: 2.6%; Multi-racial: 2.4% Econ. Disadvantaged: 42.9% Students with Disabilities: 18.5%
M6 Calculus 1 Spring	Rural Indiana Total Enrollment: 357 White: 83.5%; Black: 1.7%; Asian: 1.1%; Hispanic: 11.2%; Multi-racial: 2.2% Econ. Disadvantaged: 31.9% Students with Disabilities: 17.1%	F6 Finite Math & Calc 1 Fall (& Interview) Spring (& Interview)	Urban Indiana Total Enrollment: 3754 White: 32.5%; Black: 38.3%; Asian: 3.7%; Hispanic: 19%; Multi-racial: 6.3% Econ. Disadvantaged: 46.6% Students with Disabilities: 14.5%
M7 Finite Math Spring (& Interview)	Suburban Illinois Total Enrollment: 2688 White: 68.6%; Black: 1.6%; Asian: 6.3%; Hispanic: 19.9%; Multi-racial: 3.5% Econ. Disadvantaged: 17.8% Students with Disabilities: 14.5%	F7 Calculus 1 Spring (& Interview)	Suburban Indiana Total Enrollment: 2503 White: 77.8%; Black: 5.3%; Asian: 4.5%; Hispanic: 7.2%; Multi-racial: 4.7% Econ. Disadvantaged: 15.4% Students with Disabilities: 8.2%

Table 2. Teachers' responses to 5-point Likert-scale questions about their experiences using ALICE and questions specifically related to prompt-engineering (1=Completely disagree; 5=Completely agree).

Teachers' experiences with using ALICE	Average
Being able to work with ALICE enhanced my efficiency/productivity as a teacher	3.2
I have enjoyed this experience of using ALICE to generate questions for use in my course	4.1
Being able to create my own WeBWorK assessments felt good/empowering	3.5
I believe having ALICE as a partner/using ALICE made me more thoughtful about the content I was teaching	3.6
Working with ALICE helped me with generating creative word problems	3.8
Given a choice I would like to continue to use ALICE or a tool like ALICE that helps me with my formative (and perhaps other) assessment needs	3.9
Using ALICE required me to change how I taught math	2.5
This experience has made me aware of uses of AI that I had not thought of before	4.1
This experience has made it more likely for me to try out AI-based tools in class that help me accomplish tasks I could not otherwise	4.1
Uses of generative AI like ALICE have the potential to positively impact classroom teaching	4.4
I would recommend the use of ALICE to my colleagues	4
I am curious to learn more about integrating AI tools into my teaching or for other needs	4
Teachers' experiences with "prompt engineering"	
In general, I liked the questions generated by ALICE based on my prompts	4.2
I found that the quality of the generated assessments improved as I became more experienced with writing prompts	4.1
Over the course of using ALICE, I learned how to create better prompts for generating the desired problem	4.2
I find the process of prompt engineering to be intuitive	3.7
I believe the process of prompt engineering requires a learning curve	3.9
I found it interesting to experiment with prompts and learn from the outputs on how to create better prompts	4.1
I would have liked to get more training on how to generate good prompts	3.6
I found the process of writing prompts to get the desired question to be challenging/difficult.	2.5

Table 3.Teacher responses to the question: *"If you had to share your experience about using ALICE with teachers or future research participants, what would you say?"* and "any additional comments?"

It does a good job of giving you the types of questions that you ask, as long as you're specific enough. It's nice to have an option of generating your own questions in case you need a quick example or bell ringer.
Overall, I had a great experience. It was nice to have problems generated and automatically added into Edfinity. My only reservation I had was how long it took to get prompts back. I started submitting them once a week but then started to submit 3-4 at a time because it took so long. I felt like because I was doing that I almost forgot to submit or to add the questions to homework assignments.
Generally positive. It did what it was supposed to do. The time between prompt submission and question generation seemed way, way too long.
I would emphasize the benefits of endless generated possible questions to prevent cheating or recreating myself.
ALICE made it much easier for me to create variations of questions tgat i could use to gauge student understanding of a topic.
I think this is the same as the last open ended question. I would tell them to look at what other teachers are generating. I did not do that for a while because when I first looked there wasn't much for calculus. But, when I looked again I got some ideas on different types of questions.
Keep an open mind because it is quite easy and can save you a ton of time.
ALICE has its benefits but may not be faster than just creating your own questions.
It's a work in progress. It has the potential to be useful tool for upper-level math courses, but it still has a ways to go. Textbook companies already have online resources available that can generate ranges of values for variables, ensuring that students get unique problems each time. Being able to design your own problem using ALICE is nice, but it comes at a cost of time and patience, especially when ALICE does not generate what you were hoping for the first time.
If you want to make sure your questions are working, make sure that you have your students working in Edfinity. It was difficult to try the process out with students who work working problems in another program.
Go for it. ALICE helped create multiple problems from one prompt. It helped me see other ways to solve the problem. It was super helpful and an easy introduction to AI generated questions.
I would definitely recommend using ALICE for assessment generation to be able to create an "infinite" number of assessments to combat cheating, multiple attempts, and enhance the learning process.
This was a positive, helpful experience!
ALICE is a tool like all educational tools. Learn to use it in the right areas and it will magnify your teaching, but it will never replace the classroom teacher, nor will it do your job for you. Maintain reasonable expectations.
I used our question bank a lot while reviewing for final exams this year. I felt like it really gave my students more confidence when they went into the exam, especially since every question gave them explanations of each process.

Table 4. Responses to question about prompt engineering: *"Did you notice any patterns or strategies that consistently yielded better results when crafting prompts? If yes, please describe."*

Requesting specific things like the use of trig or the use of rational functions in a prompt.
The more specific I was, the better prompts I received.
Yes. Be more specific than you think you need to be.
I tried to generate only integers and typically -9 to 9 at times. I would have liked to try more ideas like switching with trig functions.
I felt like giving more flexibility within the prompt lended itself to better questions.
I noticed I needed to specify the types of numbers I wanted as answers so that a calculator wasn't required for solving. When I didn't do this ALICE typically gave back decimals that required a calculator. I also noticed with polynomials it did better if I guided the number of terms rather than just generically stating a polynomial.
Not that I can think of.
It doesn't need to be completely specific when writing a prompt, but also having a collecting of numbers, like "let x be a 1 digit integer" helped keep the questions from becoming too crazy
Unfortunately, the less "creative" I allowed ALICE to be, the better the results. The more I clarified every aspect of the problem, and only asked ALICE to randomize some values, the more satisfied I was with the results.
No
Using problems that I had already created for assessments were generally good ALICE prompts
Working backwards with the answer in mind, then create the parameters, then create the prompt.
Making sure I had worked out the problem beforehand to provide proper parameters / boundaries for the AI.
Yes, asking for one problem at a time, and being specific on what the outcome should look like. I would like to experiment more with various prompts and inputs in the future as well to see exactly what the limitations might be.

Table 5. Responses to the open-ended prompt "How have your views on AI changed through using ALICE?

I think it can still be really useful, especially in terms of doing tedious tasks such as writing code.
I was skeptical during the 1st semester, but I see the full benefits of how AI can help with instruction, retention, and understanding of content. (Teacher who did not sign up in the Fall term, but did in the Spring)
I found this a very useful tool as an instructor. I still worry about how students will utilize AI to their advantage to make tasks easier for them.
I believe AI is a useful tool to help create multiple variations of problems so that students cannot cheat and/or continue to practice a concept they do not understand.
I was impressed at how accurate the problems were based off of my description.
They haven't really changed a lot. I've learned more about prompting, but I had already been looking into AI quite a bit and am a big proponent for using it in education.
I didn't really go into this with positive or negative thoughts. I see that it is useful, but I also see now how it it has it's limitations. It is helpful in creating new questions, but yet getting it to understand the level of difficulty you are looking for is not easy. It has benefits, but also needs to be used in conjunction with human filtering.
It's changed significantly. I've started to be able to identify things students turn in that they had AI write for them and the ethical situations around using AI to do your work for you. I've become more accustomed to how to interact with AI to get the outcomes I need. Using AI to do things like create photos and write papers seems

unethical to me and that it should not be done. It feels like it can be used to essentially lie.
I think I understand the usefulness of generative AI much better due to this ALICE study as well as playing around with ChatGPT and Gemini on my own
I have used it more and more and feel like it is going to be very helpful in simplifying my life.
I've only had experience with AI in regards to written essays, not generated math problems from a prompt. It was interesting how I needed to change the phrasing and latex coding to create the perfect problem. I just wish it was able to created graphs and work with tables..
I understand the benefits of AI, but still worry about it in the classroom and in society.
I see AI as a tool for learning, but one that still needs to be explored. Currently it can be used to shortcut the learning process and more data needs to be collected to make sure it gets used to aid learning rather than stunt it.
I don't think I came into this with any expectations or preconceived ideas about what AI would produce which is why I stated "not sure".

Table 6. Brief case-study of 4 interviewed teachers (2 Male, 2 Female) on their overall experience, prompt engineering, how useful they found ALICE, how they used it in the classrooms, and their views on AI in education (broadly).

F5: Female, teaching Calculus 1 in rural Indiana with a mostly white population and 43% of economically disadvantaged students. **Overall experience:** Positive, found AI helpful for creating multiple problems. **Usefulness in teaching:** Specific use of ALICE problems for review and assessments, emphasizing the need for problems that are different from those in the textbook. *"I used it a ton for review... I always find when kids are wanting to review problems, I struggled with having something that's not already in their book." "Just having new resources... knowing it's available is good, is helpful."* **Student Engagement and Interaction:** The process of creating and solving ALICE problems facilitated student discussions and engagement. "It did facilitate some good discussions... Sometimes, the way I did it was better, sometimes it was a roundabout way to do it, but it did facilitate some good discussions that way." **Prompt-Engineering with ALICE**: *"It requires a learning curve... I think for me that's like, that obviously was my bad, but I think that would have helped me just because they, oh, just little language tweaks that could help.""I think as I worded it right, I feel like I got there... once I tried new question, it was okay, this didn't go what the way direction I wanted."* **AI in education**: Spoke about the dual nature of AI in education, recognizing both its potential benefits and risks of misuse.

F6: Female, teaching both Finite Math and Calc 1 at a large urban school with high numbers of Black, Hispanic, and economically disadvantaged students.
Overall Impression and reasons: Mixed (mostly negative), skeptical about practical use, prefers usual approach (existing assessments used in the past).
Usefulness: WhileF6 appreciates the ability to customize problems to some extent, there is a preference for using pre-made resources that are readily available and reliable. The need for problems to align with IU's final exams is a critical factor in F6's approach to assessment.
Other concerns include the limitations of [MLLM} such as the inability to generate pictures and the potential lowering of teachers' understanding of the subject matter. F6 believes that many teachers may not have the mathematical depth required to effectively use AI tools for creating assessments.
AI in education: F6 is cautious about integrating AI into education, citing the potential for misuse and the current technological limitations. There is a preference for traditional methods and a skepticism about the immediate practical benefits of AI tools in the classroom. Most negative (of all teachers) on "value of AI" and "trust in AI" questions in pre- and post- survey.

M1: Male, teaching Finite Math in suburban Indiana with a large Asian and Hispanic population and economically disadvantaged students.
Overall Impression and reasons: Neutral, prefers to use his bank of WeBWorK problems already created over the years in Canvas.
Usefulness: Values the ability to customize assessments to fit his specific teaching needs. Appreciates the potential of AI tools like ALICE to generate multiple versions of problems, but he also finds the current limitations frustrating. He prefers control over the content to ensure alignment with his teaching methods and standards. But also acknowledges the potential benefits of AI tools for generating practice problems and providing immediate feedback to students. He sees value in tools that can save time and enhance learning if they are reliable and easy to use.
Prompt engineering & challenges: Tried to generate problems with interesting contexts for students by having ALICE be "creative". This did not work—[MLLM} sometimes generates problems that do not align with the real-world applications suitable for students, which can be confusing for students.
The time lag in receiving feedback from ALICE and the need for back-and-forth communication to refine prompts are significant drawbacks. He mentions that this process is more time-consuming than creating problems manually, which diminishes the tool's practicality.
AI in education: A general skepticism towards new educational technologies, especially when their applicability to math is unclear. He has seen many new tools fail to deliver on their promises in the math classroom, making him cautious about integrating AI into his teaching. "*Whenever the new thing comes in professional development, they always show here's how it works in English, here's how we're socialized. Here's how it works in science. And we're pretty sure we're in math.""I know AI is the next big thing. I know, it will change. I mean, not just education, but you know, the basically the internet, like, you know, anything online."*

M7: Male, teaching Finite Math in a diverse suburb of Chicago, IL with small numbers of economically disadvantaged students. (18%)
Usefulness:
Prompt Engineering: Faced difficulties in writing clear and specific prompts initially, which required significant back-and-forth with the support team.*"There was a little bit of a learning curve... My first few problems, I probably didn't have enough parameters... But my approach was... thinking about what problem I could write that would do that."* "*The more parameters that you can offer, the better problem you can normally receive back... I needed to be... more specific with what I wanted."*
Usefulness in Teaching: M7 used ALICE to create additional practice problems, primarily as in-class practice. *"I think it would be helpful for students as well, for them to get some feedback... and be less reliant on me to... give them another problem that's like it."*
AI in Education: Generally *optimistic about the integration of AI in education, recognizing its potential to enhance teaching and learning experiences. "I find it interesting. I think it'll be helpful... I think it'll probably have a lasting place in education.*

Table 7. Teachers' pre- and post-survey responses to the question: *Describe in your own words what the term ,'artificial intelligence' (AI) means?*

Pre-Survey	**Post-Survey**
Artificial Intelligence is where a man made device makes decisions based on previously uploaded knowledge. It can adapt its decisions based on feedback.	Obtaining information or data through a computer generated system.
The computer creating content.	Artificial intelligence is a computer generating text, pictures, etc based on given information.
I would define AI as the ability of computer to reprogram itself based on new information so as to affect the way that computer performs in the future.	AI is computer implemented problem solving (broadly interpreted) that functions both in response to specific pre-programmed rules and learned behaviors acquired through training.
A computer based research of data that scans previous information from multiple sources to best answer a prompt correctly.	A computerized program that generates responses based on algorithms from millions of previous responses.
Computer generated knowledge	Computer programming that builds its knowledge through facts and procedures found on the Internet and information supplied by other users.
I usually define AI as a process or a task completed by a computer that would typically (to this point) have been done by a human because it required discernment and the ability to adapt to new information that presents itself along the way.	A computer's ability to mimic the human capacity for problem solving, learning, and creating something new.
Using computers/software to do the jobs of humans.	A simulation of human intelligence.

I don't know what AI means. I imagine i-Robot and other movies where robots take over the world.	AI is a subject with a wide scope. It can be fairly complex but it can also be using technology to adapt and change its "choices" (program choices) depending on what information is available to it.
Computers (and programs) that can "think" for themselves, creating new content and adapting to external stimuli without specific programming.	Technology that "learns" and adapts, or creates, by filtering through given information and calculating the most likely expected response.
A computer algorithm that changes as the inputs vary and the algorithm expands as more input is received. To me, it means a system of computer knowledge that can grow instead of being static.	Artificial intelligence (to me) is the application of large volumes of data across the internet to create new stuff based on what has already been made. Through probability models, AI can create similar instances of new things based on what has already been created.
Using computer generated responses to questions. These responses are not generated by a human	Using technology to solve problems and provide information without having to go through multiple sources.
In my own words, AI means computer-generated intelligence in which some type of device can perform human tasks.	Artificial intelligence (AI) is computer-generated intelligence that assists humans in tasks.
A computer or robot that is has been programmed to be able to sift through large amounts of information to have an intelligent response to a prompt	Artificial intelligence is an adaptive form of computer programming that is designed to give helpful responses to user prompts.
A machine or program designed and created by intelligent beings to replicate that same intelligence.	AI is any machine learning algorithm that learns as it iterates, gaining "intelligence" from trial and error. As it does so it begin to solve new problems and present information in a unique solution.