Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Don't Just Tell Me I'm Right: A User Study of Self-Explanation Prompting in AI-Assisted Learning

Ruby Ge

School of Computer Science and Statistics, Trinity College Dublin

Supervised by: Prof. Gaye Stephens and Dr. Sinead Impey

20 September 2026

[1] School of Nursing & Midwifery, [2] School of Computer Science and Statistics, Trinity College Dublin

Table of Contents

Abstract

Self-explanation, the practice of asking learners to articulate their own reasoning, is a well-established method to deepen understanding, yet current AI tutors rarely prompt it. This study compared two ways of eliciting self-explanation within an AI-assisted learning environment: a freeform format, where learners write their reasoning, and a menu-based format, where they select it from pre-written options. An AI tutor was built to implement both formats across a lesson on research methods. A within-subjects study involving twelve Laidlaw Scholars evaluated the strengths and weaknesses of each format, followed by a survey and interview on participants' experience and what would motivate them to engage in self-explanation within AI-assisted learning.

Freeform explanation was rated better at capturing what learners meant and more effective at supporting understanding, though it demanded greater effort; the menu-based format was less disruptive but felt more limiting. Being asked to explain an answer, rather than simply told it was correct, was perceived to deepen engagement regardless of format. These results point to recall versus recognition as the key mechanism distinguishing the two approaches, and offer a practical basis for designing AI tutors that students enjoy, return to, and learn from.

1. Introduction

One-to-one teaching can significantly improve learning outcomes compared to large-scale classroom instruction. Bloom (1984) found that the average tutored student outperformed 98% of students taught conventionally. Access to this kind of individualised instruction, however, has always been limited: a single tutor can only support a handful of learners at a time. Earlier intelligent tutoring systems(ITS) have been proven to match the effectiveness of human tutors (Nesbit et al. 2014; Kulik & Fletcher's 2016). However, this success was confined to narrow domains whose content had to be painstakingly authored by hand.

Generative AI has the potential to remove this constraint. With their human-like conversational style and knowledge drawn from vast datasets, generative AI have made immediate, personalised instruction available in almost any subject. In recent years, AI tutors have emerged as a powerful tool in education by delivering automated instruction at a scale traditional classrooms cannot match.

However, the evidence on AI's impact on learning is mixed. After all, AI chatbots are built to be helpful rather than to teach, and will prioritize delivering answers over encouraging active learning (Kestin et al., 2025). When students use AI to bypass the effort required for learning to occur, a pattern of overreliance forms, one that can lead to an erosion of critical thinking and independent problem-solving skills (Georgiou, 2025).

In practice, most interaction with an AI tutor is passive: the learner asks, and the system answers. A typical exchange asks little of the learner beyond reading. Without grounding AI tutors in real teaching practice, educational technology risks becoming disconnected from how learning actually occurs.

Self-explanation, prompting learners to articulate their own reasoning rather than simply receive an explanation, is one of the most reliable ways to shift learners from passive reception toward constructive engagement (Roy & Chi, 2005), and yet current AI tutors rarely prompt it. Self-explanation has been shown to be an effective learning strategy across

a wide range of domains, contexts, and learner populations. Its benefits in traditional learning environments are well demonstrated, but existing work replicating these findings in AI-assisted learning remains limited. Thus the question arises: How can self-explanation be meaningfully applied within AI learning tools?

2. Background

2.1 Self-Explanation Effect: Mechanisms and Evidence

Self-explanation is the act of articulating the reasoning behind a step or an answer in one's own words (Sweet, 2019). It works by prompting learners to “generate inferences about causal connections and conceptual relationships that deepen understanding”(Bisra et al., 2018). When learners self-explain, they also “detect and repair gaps between new material and their existing knowledge, which strengthens domain understanding and yields more flexible strategies for solving later problems” (VanLehn et al., 1992).

The effect was first observed in students who spontaneously self-explained physics worked examples and consequently outlearned those who did not (Chi, 1989). In subsequent experimental work, Chi (1994) showed that students who were prompted to self-explain demonstrated greater learning gains than those who were not. Since then, benefits of self-explanation have been replicated across a broad range of content areas, for learners with both high and low abilities, across domains and age groups (Rittle-Johnson, 2017), and when used with both direct instruction and discovery learning (Rittle-Johnson, 2006).

Its principal drawback is cost in both cognitive effort and time. A written explanation is effortful and time-consuming, and this burden often produces shallow responses or disengagement (Sakib et al., 2026). Aleven (2002) proposed a lighter alternative: a menu-based self explanation prompt, in which learners explain their reasoning by selecting the justifying rule from a menu rather than writing it out, which is faster and less demanding yet still improves performance on transfer problems. The two formats therefore trade depth against cost: greater effort may deepen engagement, but may equally cause fatigue or disengagement. That is the central trade-off this study aims to test.

2.2 Enjoyment and Perceived Usefulness of a Learning System

In technology-enhanced learning, enjoyment has been shown to be closely related to learning outcomes (Giannakos, 2013; Heckel & Ringeisen, 2017; Kang & Wu, 2022).

Achievement emotions such as enjoyment shape attention, motivation, and persistence, and are reliable predictors of both engagement and performance (Graesser & D'Mello, 2012).

Accordingly, this study also evaluates the affective dimensions of self-explanation on the user experience. Perceived helpfulness and enjoyment were measured using established instruments: enjoyment items adapted from the Intrinsic Motivation Inventory (Ryan, 1982), perceived usefulness items from the Technology Acceptance Model (Davis, 1989), and a mental-effort scale (Paas, 1992).

Furthermore, a tutor that helps learning but is unpleasant to use is unlikely to be adopted. So measuring enjoyment while using the instrument can also help guide how such a functionality may eventually be implemented into a commercial learning tool.

2.3 Research Goals

1. Develop an intelligent tutoring system that elicits self-explanation in both a freeform and a menu-based format.
2. Evaluate the relative strengths and weaknesses of each format through a user study.
3. Gain insight into what would motivate learners to engage in self-explanation within AI-assisted learning.

3. Methodology

The goal of this user study was to investigate how learners experience different modes of prompted self-explanation within an AI-assisted learning environment. The work was done through four stages: problem diagnosis, design and development of the learning environment, a pilot phase, and a controlled user study. This section describes each in turn.

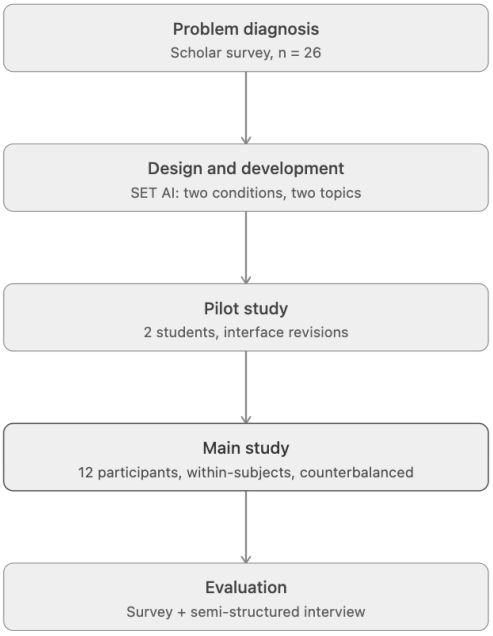

Figure 1: Stages of the study

3.1 Problem Diagnosis

A short survey of the Laidlaw Scholars network (n = 26) was conducted. It found that, despite the widespread use of self-explanation as a learning strategy, it is not something learners are often given the opportunity to do when learning with AI.

Respondents rated explaining the reasoning behind their answers as a highly useful learning technique, with a mean usefulness rating of 6.3 out of 7. However, 16 of the 26 respondents reported that AI tools never or rarely prompt them to explain their reasoning, and 20 of the 26 stated that they would value an AI tutor that prompted them to do so.

These results suggest that, at least among the population of Laidlaw Scholars, there exists a clear demand for self-explanation functionality in AI tutors.

3.2 Development of Software and Designing the Learning Environment

3.2.1 Description

The learning environment, named SET AI (Self-Explanation Tutor AI), was a custom web-based tutor built for this study. The full implementation, including source code and configuration, is available at https://github.com/99Ruby/Self-explaination-tutor

SET AI was designed to deliver lessons on research methods (See section 3.2.3 for details) by guiding them through a set of activities. Each activity presented a short scenario describing a flawed study, and asked the participant to identify the flaw and propose a corrected design, which they do with the help of an AI chatbot tutor. After the learner has shown sufficient understanding and answered the questions, they are asked to explain why their redesign works in resolving the problem in the given study. SET AI implements two explanation-prompting formats: Freeform (Chi, 1994) and menu-based (Aleven, 2002), all other elements (feedback depth, tone, interface) held constant across conditions to isolate the effect of prompt format. In the freeform condition, participants typed their explanation in their own words, whereas in the menu condition, they selected one of four pre-written options. Throughout the learning process, participants could interact with the tutor to receive hints and corrections, though the tutor was explicitly instructed never to reveal the answer directly.

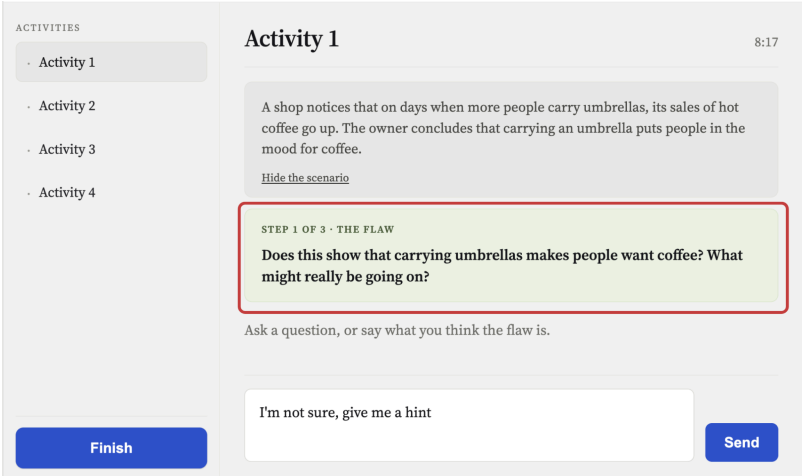

Figure 2: The learner is presented with a short scenario describing a flawed study design.

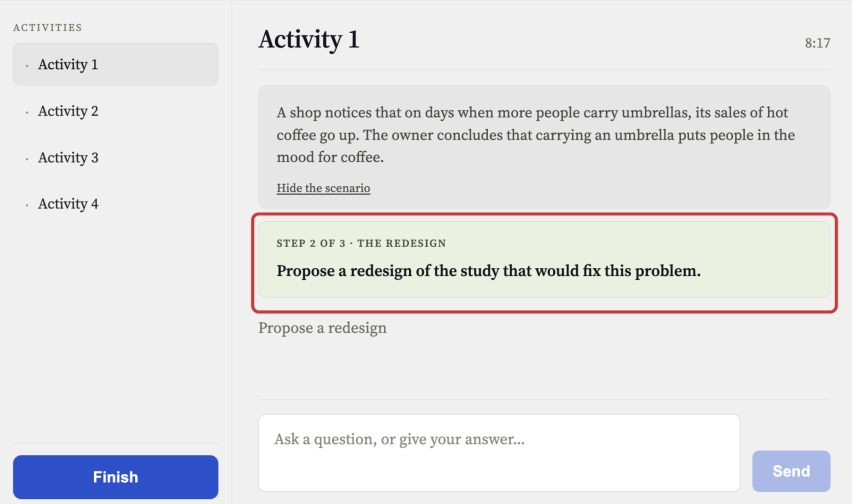

Figure 3: The learner is asked to identify what is wrong with the reasoning.

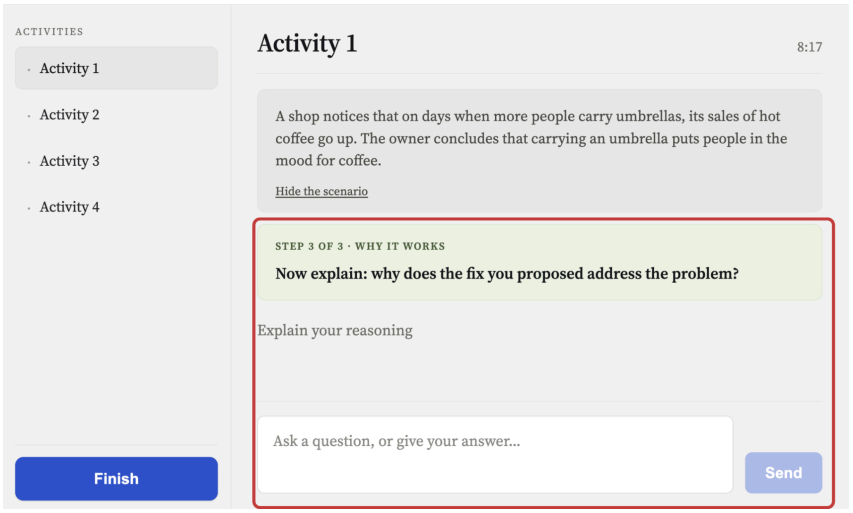

Figure 4: In the freeform condition, learners are asked to write their reasoning.

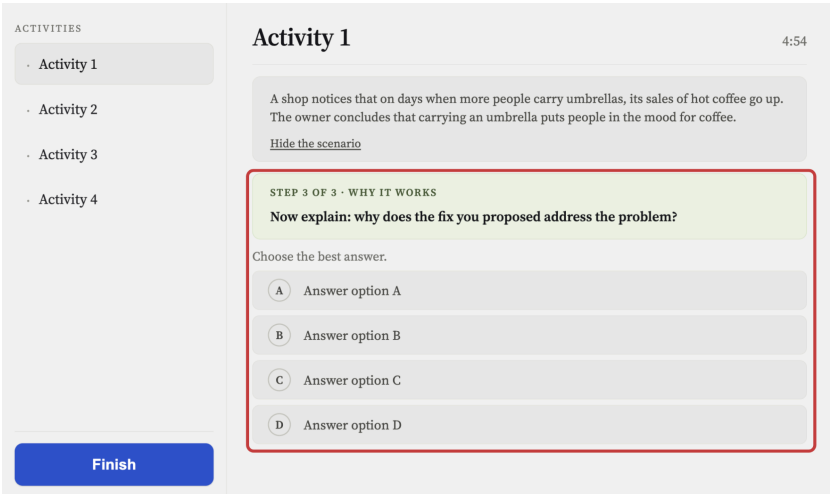

Figure 5: In the menu-based condition, learners are shown a list of explanations and asked to choose the one that best fits.

3.2.2 Tech Stack

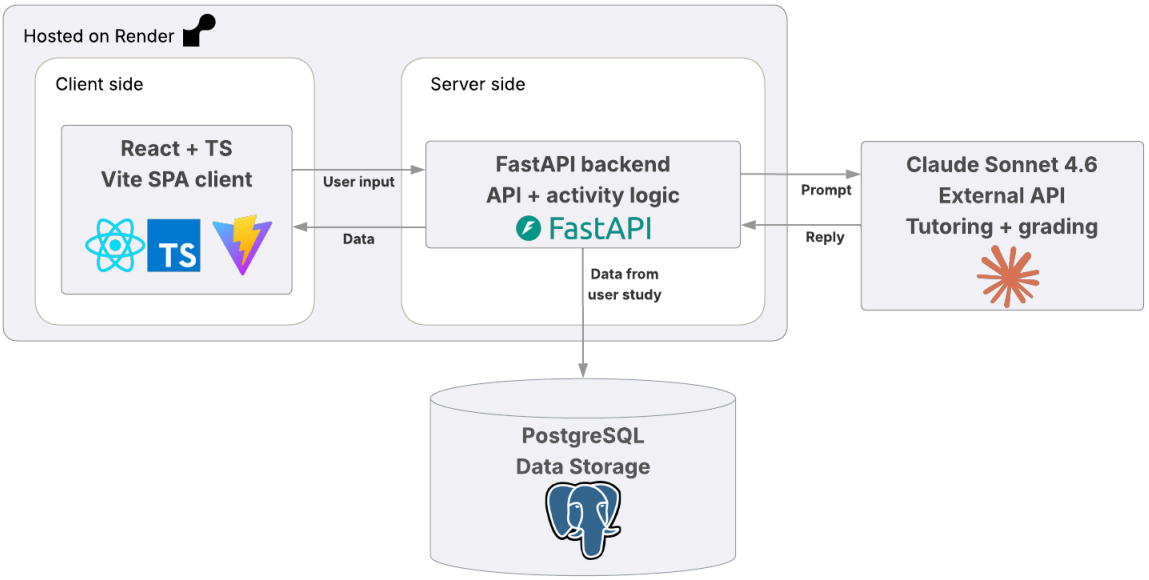

Figure 6: Tech stack of SET AI

3.2.3 Activity Content

Two sets of learning activities were created for this study. Each consisting of three “design the fix”-style questions pertaining to biases in research methods. This topic was chosen for two main reasons. Firstly, prior knowledge would be more balanced for a mixed discipline sample of undergraduate students. Any topic which draws on subject specific knowledge may lead to greater variation in how the sample interacted with the activities (e.g. a topic in mathematics may lead to STEM students finding the activities trivial). Second, research methods hold direct relevance for Laidlaw Scholars, all of whom have conducted or are currently conducting research, and so may be more motivated to engage authentically with the material.

The two lesson topics, causal inference and sampling bias, were selected to be of comparable difficulty, while remaining distinct enough to limit any carry-over effect between conditions. The full activity set can be found in Appendix A. Activity content was adapted from openly available educational sources (Streefkerk, 2023) and edited into a scenario-based format.

3.3 Pilot Study and Refinement

The interface was shaped by supervisor guidance and Nielsen's usability heuristics. Two changes were made directly from this: loading indicators were added wherever the system was fetching data, addressing visibility of system status, and button captions were reworded for clarity, addressing error prevention.

A pilot study was then conducted with two university students to evaluate the interface before the main study. Three changes were made to address the issues which emerged. First, the tutor's responses were found to be too verbose. This was addressed by adding a response length limit and adding guidelines instructing the tutor to avoid unnecessary elaboration. Second, the tutor was not guarded against prompt injection, so guardrails were added through prompt hardening. Third, the original interface separated interaction into two textboxes, one for discussion (asking questions, requesting hints, etc) and one for submitting an answer for review. This split was found to be confusing, so the two were merged into a single textbox, with a separate AI instance added to classify each input as either a discussion message or an answer attempt.

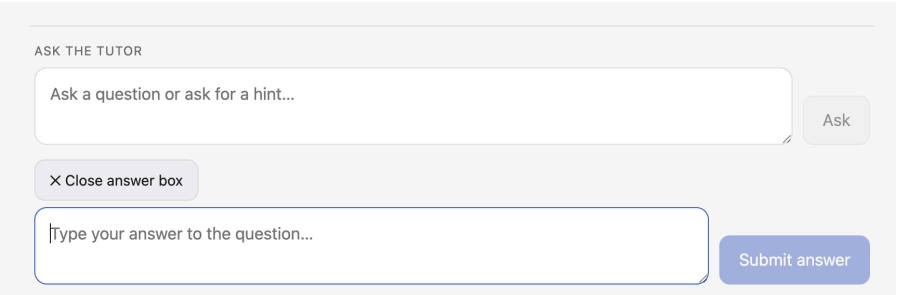

Figure 7: Dual textbox interface before pilot

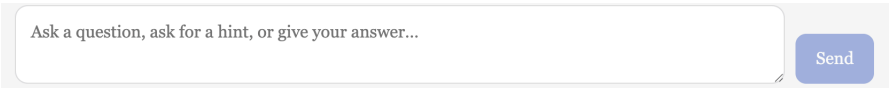

Figure 8: Single textbox interface after revision

3.4 User Study

The main study used a within-subjects design, in which each participant completed the task under both conditions (freeform and menu-based) rather than being assigned to only one.

This design was chosen because it controls for between-participant variation such as differing prior familiarity with research methods or baseline attitudes toward AI tools by using each participant as their own control. Each participant was assigned a starting combination of condition and lesson via an access code, with starting condition and topic counterbalanced using a Latin Square design, giving four total combinations distributed evenly across the sample. Participants' affective experience of the two prompting strategies was evaluated through two methods: a survey and a semi-structured interview.

3.4.1 Participants

The sample consisted of 12 current and past Laidlaw Scholars from Trinity College Dublin. Recruitment was conducted via an email sent to the Laidlaw Scholars mailing list. Participants were undergraduate students in their second, third, or fourth year of study, drawn from varied academic backgrounds including STEM, the humanities, and medicine.

3.4.2 Procedure

Sessions were conducted individually and lasted 45 minutes to an hour. After giving informed consent, each participant completed the first set of learning tasks (one topic, one condition), followed by an optional break, then the second set of tasks (the other topic, the other condition). Each set of tasks was subject to a 15-minute time limit. The researcher was present throughout but did not intervene during the tasks beyond offering technical assistance where needed. Following both tasks, a survey and a short semi-structured interview was conducted to evaluate participants' experiences of the two self-explanation prompting formats.

3.4.3 Ethics

This study received ethical approval from the School of Computer Science and Statistics research ethics board. Participants provided informed consent prior to taking part, and retained the right to withdraw at any point up until their data was anonymised. Data was stored and handled in line with Trinity College Dublin's data protection policy.

4. Results

Of the three research goals set out in Section 2.4, the first (developing a tutor capable of eliciting self-explanation in both a freeform and a menu-based format) was addressed by the design and implementation of SET AI. The results below speak to the remaining two goals: evaluating the relative strengths and weaknesses of the two formats, and identifying what would motivate learners to engage in self-explanation within AI-assisted learning. The survey results (4.1) and interview topic 1 (4.2) address the format comparison, while interview topics 2 and 3 (4.2) address learner motivation.

4.1 Survey Results

Addressing the format-comparison goal, the table below reports the mean of each survey item under freeform and menu-based conditions (1-7 scale):

Item	Freeform (1-7)	Menu-based (1-7)
I enjoyed this way of explaining my reasoning.	5.33	5.17
This way let me capture what I actually wanted to say.	5.50	3.17
Being asked to explain this way added to my understanding.	6.00	4.17
This way of explaining took a lot of effort.	2.83	1.50
The explanation prompts interrupted my flow of learning.	2.83	2.00
I felt frustrated while completing the explanation prompts.	2.67	1.83
I felt limited by this way of explaining.	3.17	4.00

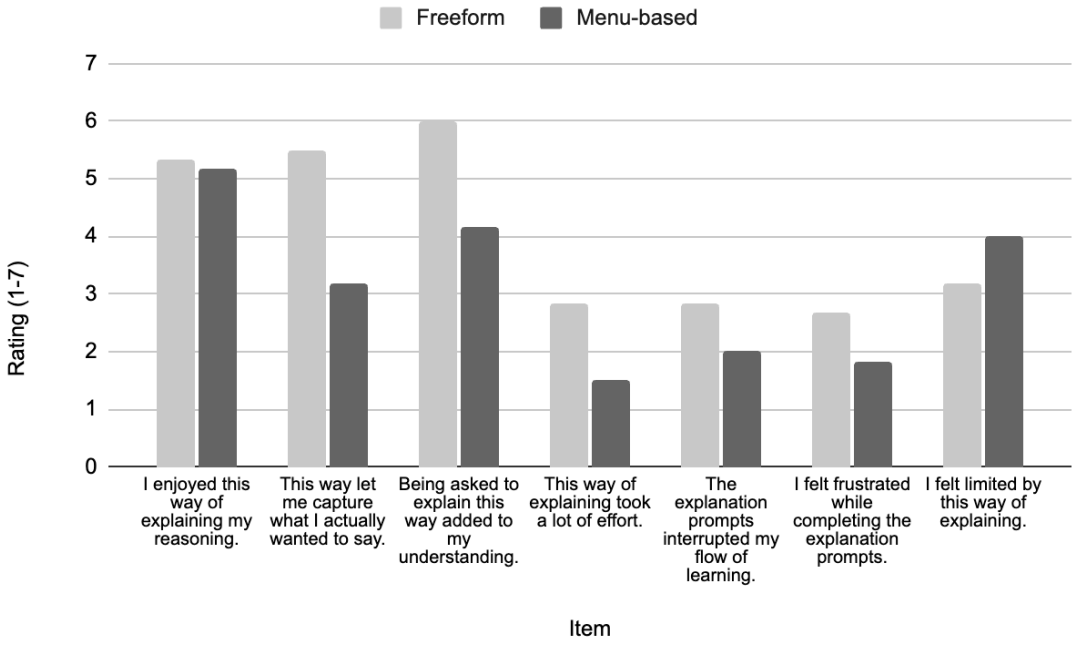

Figure 7: Survey results

Enjoyment ratings were similar between the two formats (5.33 vs. 5.17), though ratings diverged in the remaining items. Freeform explanation was rated significantly better at capturing intended meaning (5.50 vs. 3.17). Participants also rated freeform higher when it comes to contribution to understanding (6.00 vs. 4.17).

Both formats received relatively low ratings for effort taken, though freeform was slightly higher at 2.83 compared to 1.50 for menu-based. The average time spent on the explanation prompt for both formats were relatively similar: 2 minutes 45 seconds for freeform, and 2 minutes 39 seconds for menu based (the average time spent on each full task, including the explanation prompt, was 13 minutes 48 seconds). Given this near identical time investment, the effort gap likely reflects a difference in the type of cognitive work involved: the writing and active retrieval required by freeform explanation appears more effortful than the reading and recognition required by the menu-based format.

Correspondingly, menu-based prompts were rated to be less interruptive to the learning process than the freeform prompts, being rated 2.00 compared to freeform's 2.83. again plausibly attributable to the higher cognitive load involved in recall and writing rather than

to any difference in time taken. Participants also felt more limited by the menu-based format than by freeform (4.00 vs. 3.17). Frustration and disruption to flow were low in both conditions and did not differ substantially between them.

4.2 Interview Results

Topic 1: "You explained your reasoning in two ways, writing it yourself and choosing from options. How did those feel different?"

Interview responses were broadly consistent with the survey pattern. Eight of the twelve participants described freeform explanation as more engaging and more useful for learning than the menu-based format, independently using similar language. One participant noted that writing an explanation felt like using their "own words rather than putting words in your mouth," and described the menu format as superficial: that it felt like they were "picking the one that looks more correct". Another described the menu format as allowing them to "get away with doing less work," adding that this made it "less useful because [they were] thinking less," even though it was less disruptive in the moment. A third participant linked freeform explanation to better retention, describing it as "...more annoying in the moment but more conducive to retaining the information for longer". Another participant stated a direct preference for writing because it "allowed [them] to better capture [their] understanding."

Three participants expressed the opposite view, one reporting that the menu-based format "gave you a preset, defined clearly ways to think" and "felt good it gave you correct options", whereas freeform left them feeling "siloed into my own way of thinking" and "didn't feel connected to the explainer."

Topic 2: "Did being asked to explain why your answer was right change anything for you, compared with just being told you were right?"

Independently of format, most participants described being asked to explain their reasoning as creating a deeper engagement with the material. One participant noted that explaining "Helped reinforce the answer, and stopped [them] from immediately moving on to the next

thing.” Another noted that "Being told you're right feels like ‘I can do no wrong’: explaining instead produces more critical understanding." and a third that explaining forced them to think more about the answer they put down, even if it seems obvious, they had to justify "why it was obvious, like as if I was explaining to someone else." Another participant made a similar remark, that the act of explaining “makes you think more deeply. You can think you understand something, but explaining it stress-tests your understanding.” One participant took a different stance, noting that once a concept already felt understood, being asked to explain it "seems a bit pointless."

Topic 3: What would make explaining your reasoning something you wanted to do, rather than had to?

Responses to this interview topic fell broadly into two groups: internal motivation tied to personal conditions, and external motivation tied to extrinsic circumstances.
A few participants independently suggested that a learner who genuinely wanted to understand the material would already be willing to explain their reasoning and would be happier to engage in this process. Two other participants pointed to an upcoming assessment as a contributing factor: one noted they would want to explain more "if I knew I had to sit an exam after," and another gave a similar response, describing the exercise as more worthwhile "if studying for an exam."

Together, these responses suggest that motivation to self-explain depends heavily on whether the learner has a concrete external reason to engage, such as an exam, or an existing intrinsic commitment to learning the material.

5. Discussion

5.1 Linking Findings to Theory

Freeform prompting appeared to better encourage deep processing and active knowledge construction, at the cost of imposing a higher cognitive load, whereas menu-based prompting reduced extraneous cognitive load and directed attention toward the correct underlying concepts, but relied on recognition rather than recall: a mode of engagement more prone to shallow reasoning or guessing behaviour. This tracks closely with the theoretical distinction between recall and recognition memory processes, and offers a plausible account for why participants found freeform more effortful without necessarily finding it more time-consuming.

Alongside these comparisons, there was a near-unanimous view that being asked to explain, rather than simply told an answer was correct, deepened engagement with the material. This finding held independently of prompting format, and demonstrated that the act of explaining, regardless of how it is elicited, appears to interrupt the tendency to passively accept feedback and instead pushes learners to actively verify their own understanding.

5.2 Limitations

While these results are noteworthy in their own right, it's important to note that a sample size of 12 students cannot support claims about the broader population of Laidlaw scholars, let alone learners generally, and all patterns reported here should be read as provisional.

The design measures perceived experience only: no learning outcome was assessed, so no claim is made about which format actually produces more durable learning. A format learners find enjoyable or engaging in the moment does not necessarily produce the strongest retention or transfer, and testing this directly would be a natural next step.

Finally, the sample is a convenience sample of research-engaged undergraduates, which may inflate the strength of the "explaining reinforces understanding" theme relative to a more general population. As some interviewees noted: a learner who genuinely wanted to understand the material would be more willing to explain their reasoning in the first place. A

predisposition to engage in a learning process may inadvertently affect how useful they find said process.

A natural extension for future work is therefore to test whether either prompting format predicts measurable learning outcomes, which this exploratory study was not designed to assess. The relatively small sample size was appropriate given the study's exploratory aim. A future study incorporating statistical analysis of learning outcomes, however, would warrant a substantially larger sample.

6. Conclusion

This study asked how self-explanation could be meaningfully built into an AI tutor. SET AI succeeded in prompting self-explanation in both a freeform and a menu-based form, meeting the study's first goal. Across both conditions, participants described a shift from passively accepting that they were right to interrogating why. Between the two, freeform explanation better captured what students meant to say and did more for their sense of understanding, but asked for more effort in return. The menu format asked less, and gave less. Whether either was preferred came down less to the format itself than to why the learner was there in the first place, external motivation such as an upcoming exam or a wish to understand did more to make explaining feel worthwhile than interface choice could.

GenAI's promise for education is that it could make one-to-one teaching available at a scale human tutors cannot match, but a tutor that only answers, however fluently, gives up half of what made individual instruction so effective to begin with. These findings give AI tutor designers a basis for choosing between explanation formats, and a starting point for the more difficult question of what it would take to build a tutor students return to and actually learn from.

References

Aleven, V. (2002). An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, *26*(2), 147–179. https://doi.org/10.1016/s0364-0213(02)00061-7

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing Self-Explanation: a Meta-Analysis. *Educational Psychology Review*, *30*(3), 703–725. https://doi.org/10.1007/s10648-018-9434-x

Bloom, B. S. (1984). The 2 Sigma problem: the search for methods of group instruction as effective as One-to-One tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189x013006004

Chi, M. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

Chi, M. (1994). Eliciting self-explanations improves understanding,. *Cognitive Science*, *18*(3), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Georgiou, G. P. (2025). *Evidence of cognitive engagement decline*. SCALE Initiative. https://scale.stanford.edu/ai/repository/chatgpt-produces-more-lazy-thinkers-evidence-cognitive-engagement-decline

Giannakos, M. N. (2013). Exploring the video‐based learning research: A review of the literature. *British Journal of Educational Technology*, *44*(6). https://doi.org/10.1111/bjet.12070

Graesser, A. C., & D'Mello, S. (2012). Emotions during the learning of difficult material. In ˜*The œPsychology of learning and motivation/˜ The œpsychology of learning and motivation* (pp. 183–225). https://doi.org/10.1016/b978-0-12-394293-7.00005-4

Heckel, C., & Ringeisen, T. (2017). Enjoyment and boredom in academic online-learning: relations with appraisals and learning outcomes. In *Stress and Anxiety: Coping and Resilience*. Logos Berlin. https://www.researchgate.net/publication/313108939_Enjoyment_and_boredom_in_academic_online-learning_Relations_with_appraisals_and_learning_outcomes

Kang, X., & Wu, Y. (2022). Academic enjoyment, behavioral engagement, self-concept, organizational strategy and achievement in EFL setting: A multiple mediation analysis. *PLoS ONE*, *17*(4), e0267405. https://doi.org/10.1371/journal.pone.0267405

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, *15*(1), 17458. https://doi.org/10.1038/s41598-025-97652-6

Kulik, J. A., & Fletcher, J. D. (2015). Effectiveness of intelligent tutoring systems. *Review of Educational Research*, *86*(1), 42–78. https://doi.org/10.3102/0034654315581420

Michelene, T. H. (2000). Self-explaining : The dual processes of generating inference and repairing mental models. *Medical Entomology and Zoology*. https://ci.nii.ac.jp/naid/10019839747

Nesbit, J. C., Ma, W., Adesope, O. O., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901–918. https://doi.org/10.1037/a0037123

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Rittle-Johnson, B. (2006). Promoting transfer: Effects of Self-Explanation and Direct instruction. *Child Development*, *77*(1), 1–15. https://doi.org/10.1111/j.1467-8624.2006.00852.x

Rittle-Johnson, B. (2017). Developing mathematics knowledge. *Child Development Perspectives*, *11*(3), 184–190. https://doi.org/10.1111/cdep.12229

Roy, M., & Chi, M. T. H. (2005). The Self-Explanation principle in multimedia learning. In *Cambridge University Press eBooks* (pp. 271–286). https://doi.org/10.1017/cbo9780511816819.018

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, *43*(3), 450–461. https://doi.org/10.1037/0022-3514.43.3.450

Sakib, M. N., Tarin, I., Rayasam, N. M., Gaur, M., & Dey, S. (2026). ReflectED: Evaluating Reflection-Driven Learning in an AI-Assisted System. *Lecture Notes in Computer Science*, 337–346. https://doi.org/10.1007/978-3-032-29770-9_37

Streefkerk, R. (2023, June). *Internal vs. external validity | Understanding differences & threats*. Scribbr. https://www.scribbr.com/methodology/internal-vs-external-validity/

Sweet, M. (2019, December 3). *The Power of Self-Explanation | Center for Advancing Teaching and Learning Through Research*. Center for Advancing Teaching and Learning Through Research. https://learning.northeastern.edu/the-power-of-self-explanation/

VanLehn, K., Jones, R. M., & Chi, M. T. (1992). A model of the Self-Explanation effect. *Journal of the Learning Sciences*, *2*(1), 1–59. https://doi.org/10.1207/s15327809jls0201_1

Wylie, R., & Chi, M. T. H. (2014). The Self-Explanation principle in multimedia learning. In *Cambridge University Press eBooks* (pp. 413–432). https://doi.org/10.1017/cbo9781139547369.021

Appendices

Appendix A: Lesson content

Lesson 1: Causal Design					
Scenario	Question	Propose fix (free prompt)	Self-explanation (free prompt)	Self-explanation (select prompt)	Self-explanation (select prompt) Options
A company let employees choose whether to switch to a standing desk. A year later, it finds that employees who switched report noticeably less back pain than those who kept sitting desks. The company wants to publish this as evidence that standing desks reduce back pain.	As it stands, this comparison can't establish that standing desks reduce back pain. What's actually wrong with comparing these two groups?	Propose a redesign that would fix this problem.	Now explain: why does the fix you proposed address the problem with this study?	Given this fix — randomly assigning employees to a desk type instead of letting them choose — why does this address the issue of self-selection?	Correct: Because the desks are allocated by chance rather than by the employee, the two groups start out alike on average — including in how health-conscious they already are — so a difference in back pain a year later can be credited to the desk. Incorrect: Because a study in which the company decides who gets which desk is run to a professional standard, so its results carry more weight than results from a voluntary switch Incorrect: Because random assignment puts the same number of employees in each group, so neither group can outweigh the other in the comparison. Incorrect: Because employees who did not choose their own desk have no stake in the result, so they report their back pain more honestly.
A consultancy reports that companies with happier employees tend to be more profitable. It sells this to clients as: invest in employee happiness and your profits will rise.	This correlation doesn't establish that happiness drives profit. What's actually wrong with treating this correlation as evidence of that direction of effect?	Propose a redesign that would fix this problem.	Now explain: why does the fix you proposed address the problem with this study?	Given this fix — measuring happiness first and profit later, instead of observing both at the same time — why does this address the issue of reverse causation?	Correct: Because happiness is measured first and profit only later, a later rise in profit cannot be what produced the earlier happiness — the order of events now runs in one direction only. Incorrect: Because measuring the two things at different times gathers far more data than the original comparison did, so the link between happiness and profit comes through clearly. Incorrect: Because measuring happiness on its own first, before looking at profit, lets it be recorded with better tools than asking companies to describe themselves all at once. Incorrect: Because if earlier happiness and later profit rise together, a correlation that strong is enough to settle which one drives the other.

A clinic gives a new painkiller to 200 patients with chronic pain. After a month, most report that their pain has eased, and the clinic concludes the drug works.	Even though most patients improved, this doesn't show the drug works. What's actually wrong with the clinic's before-and-after comparison?	Propose a redesign that would fix this problem.	Now explain: why does the fix you proposed address the problem with this study?	Given this fix — adding a randomised, blinded placebo group instead of only tracking one group's before-and-after change — why does this address the issue of the missing comparison group?	Correct: Because the placebo group lives through the same month and the same expectation of relief, only improvement beyond what they show can be put down to the drug rather than to time or belief.Incorrect: Because splitting the patients into two groups produces two separate results, and a finding that shows up twice is more reliable than one that shows up once.Incorrect: Because patients who do not know which treatment they were given describe their pain in more precise terms, so the change is measured accurately.Incorrect: Because deciding by chance who receives the drug is fairer to the patients, since nobody is picked ahead of anyone else.

Lesson 2: Sampling Design					
Scenario	Question	Propose fix (free prompt)	Self-explanation (free prompt)	Self-explanation (select prompt)	Self-explanation (select prompt) Options
A polling company wants to estimate what share of all adults in a city support a new policy. To reach people, it dials phone numbers from the residential landline telephone directory and interviews whoever answers.	This method can't give a trustworthy estimate for all adults in the city. Whats wrong with the method?	Propose a redesign that would fix this problem.	Now explain: why does the fix you proposed address the problem with this study?	Given this fix — drawing from a frame that covers all adults, including those with no landline, instead of only the landline directory — why does this address the issue of undercoverage?	Correct: Because the frame now includes people with no landline, someone reachable only by mobile is just as able to be selected, so the result speaks for all adults rather than only landline owners.Incorrect: Because covering mobile numbers as well as landlines produces a far larger pool to call, and larger samples land closer to the truth.Incorrect: Because including mobiles guarantees that younger and older adults end up in roughly equal numbers, so neither group can dominate the result.Incorrect: Because people reached on their mobile answer in private, so they give more honest answers than people on a landline.

A council mails a survey about a proposed recycling change to a properly drawn random sample of 2,000 households. 300 households reply, and most are strongly opposed. The council concludes that residents oppose the change.	The sample was drawn well, yet the result may still mislead. What's actually still wrong?	Propose a redesign that would fix this problem.	Now explain: why does the fix you proposed address the problem with this study?	Given this fix — keeping the same random sample but working to raise and secure response instead of sampling more or different people — why does this address the issue of nonresponse bias?	Correct: Because following up the households that stayed silent brings the people with milder views into the results, so the group that answers ends up resembling the group that was sampled.Incorrect: Because reminders and follow-ups bring in more completed surveys, and a result based on more surveys is more accurate than one based on fewer.Incorrect: Because households that reply only after a reminder have had longer to think the recycling change over, so their answers are better informed.Incorrect: Because keeping the original sample rather than drawing a new one avoids surveying any household twice, so no view gets counted more than once.
A subscription app surveys its current subscribers about the service and finds that 90% are satisfied. It reports that its users are overwhelmingly happy with the service.	Surveying current subscribers can't support that claim about users in general. What's missing from who gets surveyed?	Propose a redesign that would fix this problem.	Now explain: why does the fix you proposed address the problem with this study?	Given this fix — including people who left the service in the sampling frame instead of only surveying people who are still using it — why does this address the issue of the survivorship gap in the frame?	Correct: Because people cancelled precisely because they were unhappy, adding them to the frame puts back the dissatisfied users that a current-subscriber list has already filtered out.Incorrect: Because including former subscribers makes the survey population much larger, so the satisfaction figure is estimated more precisely.Incorrect: Because people who no longer pay for the service have nothing to gain from flattering it, so their answers are more honest than current subscribers' answers.Incorrect: Because former subscribers have experience of the service across a longer period, so they are better placed to judge it than people still using it.