Coexistence of mathematical reasoning and cognitive offloading

A longitudinal case study of one student's chatbot use across a term of proof-based mathematics

Jonathan Cardoso-Silva[1,*] **and Johannes Ruf**[2]

[1]Department of Methodology, London School of Economics and Political Science
[2]Department of Mathematics, London School of Economics and Political Science
[*]Corresponding author

Correspondence
Jonathan Cardoso-Silva
j.cardoso-silva@lse.ac.uk

Compiled 23 September 2026

Abstract

Students use Generative AI chatbots for proof explanation, error checking, and exam preparation in mathematics, with documented effects on both learning and study habits. Since existing studies capture these uses in short tasks under controlled conditions, how they develop across a full academic term remains an open question. We followed one postgraduate student who studied proof-based mathematics at a research-intensive university using a Claude project configured with Anthropic's study template across an eleven-week teaching term and a revision period. We coded all 196 chat exchanges from twelve sessions based on the type of exchanges and conducted a semi-structured interview after the course. Our findings are threefold. First, the student used the chatbot productively for concept explanation, proof construction, and error correction, treating it as less intimidating than approaching teaching staff. Second, she reported spending less time thinking before turning to the chatbot as the term went on, a realisation we were able to confirm from the chatlogs later in the term. Third, the chatbot's Socratic configuration failed to match the student's needs in many exchanges, producing cascades of unanswered sub-questions, withholding worked solutions during exam revision when no solutions were available for comparison, and explaining material she already knew while skipping steps she did not understand. Our longitudinal design complements others studies by showing how these three patterns may co-exist and develop over time.

1 Introduction and Background

Generative AI (GenAI) chatbots are rapidly becoming part of students' ordinary study practices in mathematics, where a student can ask a chatbot to explain a definition, supply a missing step in a proof, check a calculation, or produce a complete solution. This availability may change not only where students obtain answers but when they decide they are stuck, how long they persist before seeking help, and how they determine whether an explanation is trustworthy. In proof-oriented mathematics these questions are especially consequential because proof comprehension, validation and construction are distinct skills (Neuhaus-Eckhardt et al., 2025), and the kind of help a student needs may depend on which of these they are trying to do. Existing studies have begun to establish that access to GenAI can improve performance while using the tool without necessarily producing corresponding gains in independent performance (Bastani et al., 2025; Rismanchian et al., 2026), and other work has examined how students regulate their learning and offload cognitive work during relatively short tasks designed by researchers (Fan et al., 2025; López-Pernas et al., 2025). These studies reveal comparatively little about how students incorporate general-purpose GenAI tools into their ordinary study practices over time, and in particular about how a student's relationship with the tool changes through successful and unsuccessful interactions across a full teaching term.

This paper investigates how one postgraduate student (pseudonym Lena) used Claude, a general-purpose GenAI chatbot, as a study tool across a proof-oriented mathematics module at a research-intensive university. The student configured the chatbot using Anthropic's study template, which gives it a Socratic stance, and used it across eleven teaching weeks and a revision period, producing twelve chat sessions containing 196 exchanges. After the teaching and assessment period, a semi-structured interview captured the student's account of her study practices, her use of the chatbot, and how she believed her approach to studying had developed over the term. The interaction record allows us to observe how the student requested assistance, responded to questions and explanations, challenged errors, and

sometimes left an exchange unresolved, while the interview provides her retrospective account of why she used the tool and how she interpreted its behaviour.

After reviewing work on mathematical help-seeking, Socratic chatbot modes and cognitive offloading, we examine how the student's use changed across the term and how the usefulness of the chatbot's support depended on both what she already understood and the mathematical activity she was trying to undertake.

Help-seeking behaviour in proof-based mathematics

Well before GenAI, studies of mathematics students distinguished instrumental help-seeking, where a student asks for enough assistance to continue working independently, from executive help-seeking, where the student asks for the answer itself. Aguilar and Puga (2020) observed both behaviours among undergraduates using the internet for calculus: students worked through procedural tasks with tutorial material but copied definitions for declarative ones.

Proof validation creates a related difficulty. Students have been found to judge proofs by whether the conclusion is correct or agrees with their own answer rather than by reconstructing the reasoning and warrants connecting each step (see Yoon et al., 2024). Similar patterns appear with GenAI. Zhuang (2026) found that calculus students commonly judged ChatGPT-generated solutions by whether they were correct, understandable, or matched their own work; only one checked whether the conditions of the relevant theorem were satisfied. In proof tasks, students have likewise accepted AI-generated arguments without examining their justification or appropriateness (Dilling and Herrmann, 2024; Klawa et al., 2025; Yoon et al., 2026; Urhan and Zhuang, 2026). These studies have largely examined researcher-designed tasks or structured sessions. Our case asks how such help-seeking and verification develop when a student uses a chatbot independently across a full teaching term.

The uncertain pedagogical value of Socratic configurations

By 2025 and early 2026, major commercial GenAI chatbots had introduced "Socratic" study modes that favour questions, hints, and staged assistance over immediately providing a complete answer.[1] Evidence for their pedagogical advantage is mixed. Bastani et al. (2025) found that high-school mathematics students using a generic chatbot performed worse on a subsequent exam than a control group, while those using a Socratic version performed about the same as the control group. Other studies have likewise found benefits while using guided chatbots without clear gains on later independent tasks (Makransky et al., 2025; Jin et al., 2026).

These modes rely heavily on instructions that tell the model how to respond. Prompt-level guardrails can encourage questioning, but do not by themselves provide a sufficiently detailed or current account of what a student knows or of the mathematical activity they are trying to undertake (Khosravi et al., 2026; Schell et al., 2025). Effective questioning, by contrast, depends on responding to the student's developing understanding (Zhuang and Conner, 2022). Students can partly compensate by asking better follow-up questions themselves (Zhuang, 2026). Our case study traces how these tensions played out when one student used a Socratic configuration across a full term of proof-based coursework.

Cognitive offloading and mathematical verification

Cognitive offloading is the use of an external resource to reduce the cognitive demands of a task (Risko and Gilbert, 2016). With GenAI, students can offload parts of the mathematical reasoning a task is intended to develop. What matters is what work is delegated: asking a chatbot to supply a proof step may remove the need to construct that step while still leaving the student to understand or validate it.

Experimental evidence suggests that this distinction matters. Bastani et al. (2025) found no measurable harm on a later mathematics exam when students used a chatbot that offered hints rather than answers, while Abdelghani et al. (2026) found better later performance among students who continued asking conceptual and procedural questions than among those who increasingly asked for verification or answers. Ready access to AI can also reduce independent regulation, a pattern Fan et al. (2025) call "metacognitive laziness."

Verification becomes particularly important when the chatbot itself is wrong. In one mathematics study, ChatGPT agreed with students' critiques and replaced previously correct solutions with incorrect ones (Zhuang, 2026). GenAI

[1] Anthropic introduced learning mode with Claude for Education in April 2025 (https://www.anthropic.com/news/claude-for-education), OpenAI launched ChatGPT study mode in July 2025 (https://openai.com/index/chatgpt-study-mode), and Google added guided learning to Gemini in August 2025 (https://gemini.google/is/release-notes).

use may also change over time: Misiejuk et al. (2026) found that prompting patterns diverged across successive coursework submissions, although the data could not distinguish growing skill from growing reliance.

2 Research design and methods

Research design

The studies reviewed above draw on controlled tasks, structured lab sessions, or aggregate course-level data and have not followed a student's own use of a chatbot across a full teaching term in proof-based mathematics. It is therefore not known whether the patterns observed under controlled conditions persist, develop, or reverse when students study on their own over weeks and months. This study addressed three questions:

RQ1. How did the student's study practices with the chatbot develop across the teaching term?

RQ2. When did the chatbot's Socratic configuration support the student's mathematical work, and when did it obstruct it?

RQ3. How does the student interpret her own experience of studying with the chatbot?

We conducted an instrumental case study (Denzin and Lincoln, 2018, chap. 14) of one student who used a GenAI chatbot while studying a proof-oriented postgraduate mathematics module at a research-intensive university in the United Kingdom. The module covered mathematical definitions, proofs, derivations, algorithms and asymptotic bounds. At the start of the teaching term, the first author visited the class of eight students, explained the study, and invited them to share their GenAI chatlogs. Students were free to use any chatbot they wished, and no instructions were given on when or how to use it. The first author sent weekly emails requesting logs. The second author taught the module, and the first author, who is a lecturer in a different department at the same institution, had no teaching role in it and no prior contact with the participant.

One student agreed to participate. The participant (pseudonym Lena) was enrolled in a Master's programme in operations research and was taking the module as an optional course whose subject matter was largely new to her. She produced twelve chat sessions containing 196 exchanges between 22 January and 6 May 2026, spanning the eleven-week teaching term, including a reading week at Week 6, and the revision period that followed the final lecture in early April. After the teaching and assessment period, the first author conducted a semi-structured interview in which the participant described her study practices, her use of the chatbot, and how she believed her approach to studying had developed over the term.

Case studies have long been used in mathematics education to provide detailed, context-dependent accounts of students' mathematical activity (Iannone et al., 2020; Locke et al., 2023). Existing studies of GenAI and mathematical proving have also used small samples, but largely in controlled or short-term settings (Yoon et al., 2024; Zhuang, 2025, 2026; Urhan and Zhuang, 2026; Klawa et al., 2025). Our case instead follows one student's self-directed use across a full teaching term, allowing us to trace how the same person used the same tool differently across mathematical tasks and stages of the course (Flyvbjerg, 2006).

Student's configuration of the GenAI tool

The participant used Anthropic's Claude, a general-purpose GenAI chatbot, accessed at no cost through her institution's Enterprise subscription, and it was the only GenAI tool she used for the module. Claude allows users to create *projects*, persistent workspaces whose instructions set the chatbot's behaviour across all conversations within them. The participant created a project using Anthropic's study template, a set of default instructions that give the chatbot a *Socratic stance*: it favours guiding questions and staged assistance over immediately supplying a complete answer.[2] The participant's only deliberate change to these instructions was to request that mathematical notation be formatted in LaTeX (see Supplementary Material A for the full project instructions). The participant also uploaded the course lecture notes to the project. Claude projects include a memory feature, enabled by default, that builds a daily summary of the user and their interactions across conversations. Every time the participant started a new chat within the project, the

[2] See https://info.lse.ac.uk/current-students/digital-skills-lab/claude for the institution's Claude for Education provision, and https://www.anthropic.com/solutions/education for Anthropic's description of learning mode and project templates.

chatbot could read the project instructions, the uploaded lecture notes and this accumulated summary, and it drew on them when composing its replies from the first message, though it could not see what had been said in any previous conversation. Note that the participant could ask for a direct explanation or solution at any point, and not every exchange followed the questioning pattern.

Data sources

The main data are the twelve project chatlogs (196 exchanges) and a semi-structured interview conducted after the teaching and assessment period; course materials and the project configuration were used to interpret the mathematical content and the chatbot's behaviour (Table 1). We call each student message and the response that followed it an *exchange*, and consecutive exchanges concerned with the same mathematical question or study purpose an *episode*. Exchanges are numbered c0–c195, and the twelve chat sessions were kept separate because each new conversation began without the previous conversations themselves.

Table 1: Data inventory for the case study.

Source	Extent	Period
GenAI chatlogs	12 chat sessions, 196 exchanges	22 January – 6 May 2026
Semi-structured interview	1 audio recording, approx. 50 min; 1 transcript	8 July 2026
Course materials	9 sets of lecture notes with slides 9 problem sets with solutions 2 exam papers	January – May 2026
GenAI project configuration	1 project instruction file	January – May 2026

The interview was informed by prior reading of the chatlogs and organised chronologically, following McGrath et al. (2019). Particular episodes were introduced only after the associated topic had arisen in the participant's account, so that the interview was not structured around judgments already made by the researchers. The first author emphasised that there were no correct answers, avoided evaluating the participant's responses, and used paraphrasing and follow-up questions to check her interpretation.

Ethics approval. Ethical approval for the project was sought and granted under the London School of Economics and Political Science's process for institutional research, analysis and evaluation (IRAE ref. 212088, approved February 2026). The participant gave informed consent for the researchers to analyse her chatlogs and interview transcript and to reproduce de-identified excerpts in publication. Participation was voluntary and had no effect on teaching, assessment or progression.

Analysis

The analysis proceeded in four steps:

1. Corpus construction and familiarisation. The first author reconstructed the course timeline and grouped exchanges into episodes. Both authors then read the chatlogs together in chronological order. The second author contributed judgments of the mathematical content and course context, while the first author recorded the mathematical task, the student's apparent purpose, visible evidence of prior work, and the subsequent direction of the interaction.

2. Coding each exchange. For each exchange, the first author coded the student's prompt type and the chatbot's response type, and recorded the mathematical task, visible prior reasoning, mathematical and curricular adequacy, uptake, verification, and resolution. Correctness was distinguished from pedagogical adequacy. The coding drew on the GENIAL framework (Cardoso-Silva et al., 2025), with additional codes introduced for recurring features of the mathematical interactions. The coding scheme, with definitions, inclusion criteria and worked examples, is provided as Supplementary Material B.

3. Analysis of the interview. The first author reviewed the transcript for the participant's reported purposes, study practices, perceptions of the chatbot, and perceived changes over time, and connected these to episodes where her account matched, extended, or contradicted the chatlogs.

4. Identifying themes. The first author compared episodes across the term, particularly cases where Socratic questioning supported or disrupted mathematical activity and where offloading coexisted with close reading, correction and verification. The authors revised the resulting themes together, discussing the mathematical content and course context.

Use of GenAI in the analysis

At several points during the analysis, the first author used Claude Opus 4.6, accessed through the university's Enterprise subscription, to assist with chronological annotation and coding of the chatlogs, linking interview material to particular exchanges, and reviewing candidate themes. In each case, the interpretive work described above preceded the model's use. Where its codes or suggested connections differed from our own, the first author checked them against the codebook, chatlogs and interview transcript; outputs that flattened contradictions or misrepresented the participant's perspective were discarded. The first author wrote all substantive text and decided which themes and claims entered the paper, and no model-generated claim was retained without verification against the underlying data.

Each use of the model was run in a separate session with its own instructions. The full workflow instructions and illustrative outputs are provided in Supplementary Material C, following Kempny et al. (2026). Our use of the model treated it as an analytical aid rather than an independent interpreter, consistent with calls for technological reflexivity in qualitative GenAI research (Ibrahim and Voyer, 2026). The second author contributed judgments about the mathematics and course context, while the first author led the qualitative analysis. Identifying details were removed or generalised where they were not needed for the analysis.

3 Findings

How the student studied

The student described her approach to the course as reading the lecture notes, summarising them, and tracing every step the lecturer had taken, turning to Claude if she got stuck on a step or wanted to check her own working. Table 2 shows how these exchanges were distributed across the teaching term. The student used the chatbot most heavily in the first two weeks of the course, when the material was new to her, and again after the final lecture, when she was preparing for the exam.

In several weeks the exchanges concerned material from earlier lectures, not the week's scheduled topic. The student did not use Claude for this course in W03, W06, W07 or W11. By the revision period the student had stopped using Claude for new material and fell back on traditional study: repeating the problem sets and writing flashcards. She still turned to Claude during this period, but for drilling past-exam questions and checking whether topics were still on the syllabus, and the tool performed poorly on both (see the case note on exam preparation below).

Overview of the interactions

Most of the student's exchanges with Claude asked it to explain a concept or a step in a proof (80 of 196). Claude answered directly in 60 of these exchanges and posed guiding questions in 15. The second most frequent prompt type was a response to a Socratic question Claude had posed (57 exchanges). The four most common chatbot responses to these prompts were further questions (24), corrections (12), direct explanations (12), and confirmations (8). The student rarely asked how to approach a problem without showing her own attempt first (6 exchanges, all but one in a single session on 6 May, four weeks before the exam). Table 3 summarises the full distribution.

Patterns in how the student used the chatbot

Beyond these numbers, three major themes stood out to us while looking closer at the exchanges, interpreting the intent behind each query Lena posed to the chatbot, and taking into account the observations she made about her own usage.

Table 2: Distribution of chat exchanges across the teaching term.

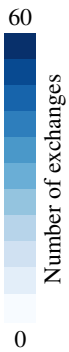
Week	Scheduled lecture	Dates	Exchanges	*n*	Material studied	Chatlogs[a]
W01	No-regret algorithms	19 Jan – 25 Jan	c000–c036	37	W01–W02	1[b]
W02	Mechanism design basics	26 Jan – 01 Feb	c037–c046	10	W02–W03	1[b]
W03	Optimal mechanism design	02 Feb – 08 Feb	—	0		0
W04	Approximate mechanism design	09 Feb – 15 Feb	c047–c054	8	W02	1
W05	Contract theory	16 Feb – 22 Feb	c055–c066	12	W03, W05	2
W06	Reading week	23 Feb – 01 Mar	—	0		0
W07	Randomised algorithms	02 Mar – 08 Mar	—	0		0
W08	Streaming algorithms	09 Mar – 15 Mar	c067–c081	15	W07–W08	1
W09	Frequency moments	16 Mar – 22 Mar	c082–c098	17	W09	1
W10	Random structures	23 Mar – 29 Mar	c099–c128	30	W09–W10	2
W11	Exam preparation	30 Mar – 05 Apr	—	0		0
	Revision period	13 Apr – 06 May	c129–c195	67	W02–W09	3
	Exam	02 Jun 2026				
	Total			**196**		**12**

[a]Each chatlog is an independent conversation; the chatbot could not see what had been said in previous chatlogs, though it had access to a memory summary that accumulated across conversations (see §2). [b]One chatlog (38 exchanges) straddled the W01/W02 boundary.

Table 3: Student prompt type by chatbot response type ($n = 196$), first author's primary codes. Colour represents count of combinations between `prompt_type` and `response_type` codes. Code definitions are provided in Supplementary Material B.

Student prompt type (first author's code)	EXPLANATION	CODE	REDIRECT	CORRECTION	CLARIFICATION	AFFIRMATION	*n*
CHECK	3	2	5	5	0	1	**16**
REFINE	8	1	0	0	1	0	**10**
EXPLAIN	**60**	0	15	3	2	0	**80**
CONCEPT	6	0	2	0	0	0	**8**
META	4	1	0	0	0	0	**5**
SOLUTION	3	2	2	0	0	0	**7**
RESPOND	12	0	**24**	12	1	8	**57**
OTHER	2	4	1	0	0	0	**7**
HOWTO	2	0	4	0	0	0	**6**
n	**100**	**10**	**53**	**20**	**4**	**9**	**196**

Chatbot response type (first author's code)



The less intimidating explainer

In 80 of the 196 exchanges, Lena asked Claude to explain a concept or a step in a proof. From her first interactions with the chatbot she sent expressions she had derived and asked how to get from them to the result shown in the lecture (c3–c5), and in the interview she described this as a replacement for contacting teaching staff: *"For me, that's definitely something that the AI is replacing, like literally going there and asking the professor."*

She felt some of her questions were too basic to take to the lecturer because the topic was new to her. Claude *"took away a lot of that pressure of going somewhere and asking the professor a question that you should."*

Her use of the chatbot was not simply passive answer-seeking. In one exchange from the first week (Figure 1, c10), she asked how anyone would ever think of a bounding step used in a proof: *"I understand that it's true but I wouldn't get it myself."* She wanted to understand the mathematical reasoning that produced the step. In another exchange, where Claude posed its typical Socratic guiding questions (c72–c73, Figure 2), Lena disagreed with its reasoning about a streaming algorithm, constructed her own counterexample, and held on to it through several exchanges, working through the counterexample with Claude.

STUDENT 17:37:08

I understand that its true but I wouldn't get it myself

CHATBOT 17:37:31

This is definitely not something you'd be expected to derive on your own during an exam. Let me explain where it comes from and how you might think about it. ...

Method 1: Taylor Series (Most Common Derivation). The Taylor series expansion of $\ln(1-z)$ around $z=0$ is ...

Figure 1. Exchange c10, 22 January 2026. The student asks how anyone would arrive at one step in a proof from the Week 1 lecture. The reply gives two derivations and reassures her that the step is beyond what she would need to reproduce.

STUDENT 10:10:11

I'm confused on the finding a duplicate (multi-pass). can you explain to me what it does? since its not ordered, can't the duplicates just be inconveniently placed so we never find them?

CHATBOT 10:10:11

Let me walk you through the key insight. The algorithm isn't searching for duplicates by looking at pairs of elements — it's using a pigeonhole-style argument on *counts*. ... So what can you say about at least one of those halves — how many stream elements must fall into it compared to how many distinct values that half contains?

STUDENT 10:11:47

BUT the smaller half could have the duplicate, e.g. 22341 with m=n+1 and n = 4

CHATBOT 10:11:47

But here's the key point: the algorithm always follows the "overfull" half, because that's where a duplicate is guaranteed to exist. ... Now, what if instead the stream were 2 2 3 4 3?

Figure 2. Exchanges c72–c73, 12 March 2026. The student asks about an algorithm from the streaming-algorithms lecture. Highlighted text marks where the chatbot poses a Socratic question. In the next exchange (c74, not shown) the student still splits by position, and the response walks the full procedure on her numbers. Compare Figure 4, where the questions expand and the student gives up.

In both exchanges Lena retained control over her own understanding and used the chatbot to test and extend it. Other interactions were less successful, however, and she described some as *"wasting my time,"* a question we return to in How well did the chatbot respond to the student? below.

Doing the mathematics vs asking for it

Over the term, Lena reported spending less time thinking before turning to Claude. She put it this way:

> *My frustration level has sunk, like the accepted time I'll think about something before I ask Claude, like, how do I start thinking about this? Before, I just had to do it by myself, and now I really quickly turn to Claude and be like, okay, just give me a point of where to start.*

The chatlogs are consistent with a change in what Lena brought to Claude, although they cannot show how much independent thinking occurred before a message was sent. In January, at the start of the course, she would derive an expression and bring it to Claude for checking (c3), and in one instance worked a derivation across five exchanges before Claude's output suggested the approach would not work (c7). By Week 4 she was sometimes asking in one line, with no working shown (c50) – see Figure 3. By Week 10 she pasted the title of a lecture slide about an estimator and asked Claude to explain the whole algorithm, again without showing any working of her own first (c103).

STUDENT 16:48:40

for the next sentence "A similar argument shows that for $n \geq 2$ the external regret achievable by any algorithm will be at least $\Theta(\sqrt{T \ln n})$", I got this up until the point where I have to find that $\sqrt{T/n^n}$ is $\Theta(\sqrt{T \ln n})$ — how would I get that?

CHATBOT 16:48:52

Great question! Let me help you see where the $\ln n$ term comes from. It sounds like you're on the right track, but let me clarify the argument ...

(a) Exchange c3, 22 January 2026.

STUDENT 19:58:44

why does the Proposition 2.19. automatically follow

CHATBOT 19:59:01

Good question! Let's work through this. Proposition 2.19 says: If $M = (x, p)$ is DSIC, then x must be weakly monotone ...

(b) Exchange c50, 15 February 2026.

Figure 3: Two exchanges three weeks apart. In (a) the student brings her own derivation and asks about a step she is stuck on. In (b) the question is one line, with no working shown. The contrast in what the student shows is clear. Whether it reflects less independent thinking or only a shorter message is not settled by these two exchanges alone. Both responses open the same way. Mathematics has been typeset for readability; the student sent plain text, not LaTeX.

Lena also said that she did not always read Claude's full responses. In one exchange she asked for an inductive hypothesis that the previous response had already stated (c23, c115), and in several others she moved on before engaging with material beyond what her question required (c11, c40, c57, c64, c116).

At the same time, she continued to engage closely with the mathematics. In the same period she caught a missing step in a Jensen inequality derivation (c111), spotted a missing case in a proof (c101), and during revision questioned a term in a derivation Claude had produced (c134). Offloading and close mathematical engagement therefore coexisted across the same sessions and sometimes the same afternoon.

Self-correction, trust, and the limits of verification

Lena described her process with Claude as trying to *"kind of understand where it's coming from, what it's saying, not just copy-pasting it,"* and then deciding whether the answer made sense. She caught errors in Claude's mathematics seven times across the term. These included wrong arithmetic that she corrected herself (c5, c29), a missed case that she defended when Claude initially dismissed it (c34, c35), a skipped step in a Jensen inequality derivation (c111), and an unstated use of Stirling's approximation whose claimed result she checked against Claude's own method (c120, c121). During revision, after an eighteen-minute gap, she returned to reject a proposed counterexample: "this one doesnt work" (c175).

Claude sometimes revised its own derivations mid-response, writing "no actually" or "wait, let me recalculate" before starting again; we observed this three times without prompting from Lena (c4, c7, c15). She interpreted this as a form of checking, although her trust remained ambivalent:

> *When it does that, in some way it makes you feel more secure because you know it's cross-checking, but then also you don't trust the answer as much because this one might also be wrong.*

These rewrites were not formal verification, despite Lena interpreting them as evidence that Claude was checking its working.

Her ability to verify Claude's mathematics depended on what she already knew. MA421 was her first proof-based course: she had learned induction as an undergraduate but had *"never read a mathematical proof before or . . . done one myself"*. She could check arithmetic, but an incorrect step in an unfamiliar proof was harder to recognise. At times she therefore asked Claude itself to locate the relevant result in the lecture notes (*"give me the point in the lecture notes where I would cross-check this"*), relying on the tool both to supply a missing step and to help her locate the material against which to verify it.

How well did the chatbot respond to the student?

In this section we turn our attention to the chatbot's responses and to how the project's Socratic configuration shaped what she received. We will see how that setup gave her a capable study tool for working through proofs, but also how the same Socratic configuration created difficulties in other exchanges.

What the chatbot did well

On a Chebyshev exercise in Week 9, Lena could not think of how to get started and asked Claude for help. Claude defined the random variable and the goal, then asked which inequality she would apply (c90). She identified the correct inequality, and over the next four exchanges the chatbot caught an error in her bound and walked through the remaining steps once she had gone as far as she could (c91–c95). Here Lena had made clear what she did and did not understand, and the chatbot matched the depth of its support accordingly.

Incorrect mathematics and uncritical agreement

The chatbot also produced incorrect mathematics. While revising for the final exam, Lena asked it to explain a streaming algorithm from a past paper, and Claude incorrectly claimed that the line $s \leftarrow x_i$ was a typo for $s \leftarrow p(x_i)$ (c158). Lena disagreed, but Claude maintained the erroneous claim across three further exchanges (c159–c161).

The opposite problem also occurred. Lena told Claude that a definition it had cited was not in the lecture notes, and the chatbot apologised and agreed (c056), even though the definition appears on the second page of that lecture. In both cases, assessing the response required knowledge independent of the chatbot itself.

How Socratic questioning worked against the student

In several exchanges Lena came with her own work already done or was stuck on one specific derivation step, but the Socratic configuration opened with guiding questions before giving direct help (c33, c45, c58, c59, c133, c139). She described how these could multiply: *"It gives me five really great conceptual questions that I should think about, but then it gets so lost in the tiny little sub-questions that I don't even get to talk about the five big ones."*

The most striking examples came from the first problem set. Lena explicitly asked Claude to guide her rather than provide solutions (c17). On one exercise, 28 questions accumulated across eight exchanges; on the next, Claude posed 33 across twelve exchanges, of which Lena answered four while the rest accumulated unanswered or were re-asked. She never completed the second exercise (Figure 4). Similar patterns appeared later (c36, c41, c103), and in one case she gave up within a minute and wrote *"I don't know, please tell me"* (c82, c83).

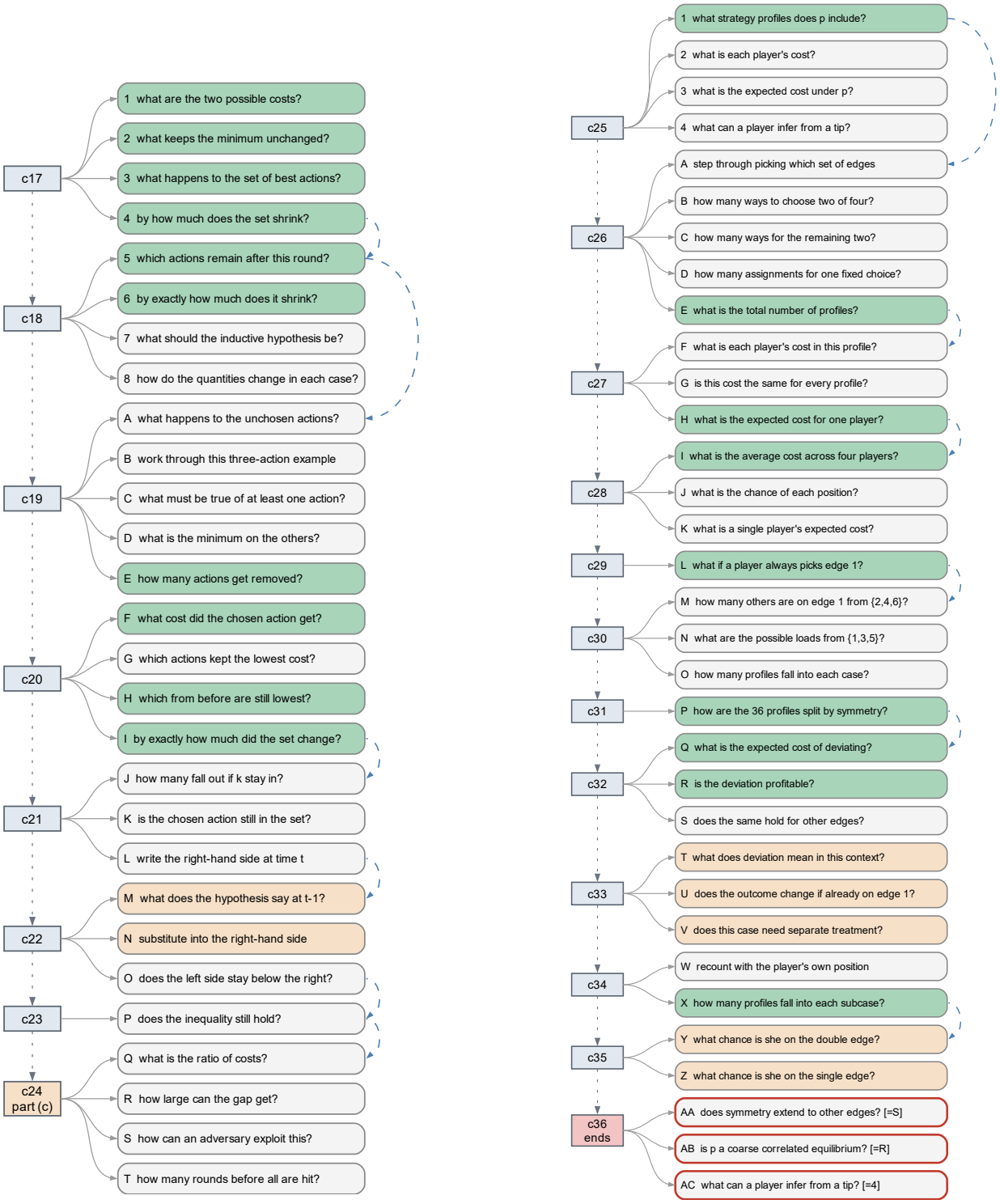
Other mismatches were less extreme. Claude sometimes repeated material Lena had asked it to simplify (c109), supplied tangential examples (c13), or added exam advice when she had asked about a proof step (c10).

What the student wished the chatbot did differently

In the interview, the student was asked what would have made the chatbot more useful:

> *Maybe choosing like the level of where I want to enter the conversation. Like if I wanted to explain it from scratch, like all the basic concepts in detail, or if I can be like, oh, I've understood most of it. I just have like a few specific questions.*

By "level" she referred not simply to the difficulty of the mathematics but to how much she already understood and the form of support she wanted. Her distinction was between being walked through a proof step by step, having her



(a) Exercise 1.1(b) and part (c). 28 questions, 8 exchanges.

(b) Exercise 1.2. 33 questions, 12 exchanges.

Figure 4: Every question Claude posed on Problem Set 1 (Lecture 1). Answered: the student answered this question. Unanswered: posed but never answered. Self-answered: the response posed and answered it itself. Re-asked: dropped, then re-asked in the final message. Blue dashed arrows connect each answered question to the follow-up it produced in the next exchange.

own work checked, or receiving a summary connected to what she already knew. She thought better prompting on her part could have achieved this: *"I could have prompted that, that was on me"* and *"maybe I should have prompted it in those specific chats more specifically, just like give me the solution directly, because sometimes you just want to cross-check what you did."* Asked whether the chatbot's setup should account for this without better prompting, she suggested *"an initial exam to like kind of know where you are"* and settled on level of knowledge as the thing she would want to set.

Every now and then, the student did specify her level before asking and the chatbot complied. In one instance, she wrote which equation she could follow and which she could not (c12), and the response matched the depth of its answer accordingly. The very next exercise, however, produced the ever-expanding questioning shown in Figure 4. Later, during the revision period, after the Claude project had accumulated a memory summary across earlier sessions, a fresh session opened with Socratic questions pitched below the student's level on a problem set exercise she could solve in under a minute (c129). During exam revision, she had a past paper without solutions and asked the chatbot to produce them. The chatbot's Socratic stance did not provide them, instead opening with unrelated guiding questions (c165). A direct answer, or scaffolding that built on her existing knowledge, would probably have worked better. As no solutions to that paper existed, she had no answer key against which to verify her own answers.

Prepping for the exam

The student kept using Claude through the revision period, but the task changed from understanding new material to drilling what she already knew, and the patterns described in the three themes above followed her into it.

The module had no past papers that matched the current syllabus, because the course changes topic coverage each year. The student had access to an old exam and used Claude for three things: working through solutions to its questions (c156–c164), checking which questions were still relevant to the current curriculum (c180, c195), and generating new practice questions. When working through solutions, the responses contradicted themselves and did not produce a correct answer (c164, c178). On the relevance check, the student said:

> *"It basically find questions as relevant that are definitely not relevant. But then Claude told me, so I was like, oh, maybe, maybe they are relevant. [It] freaked me out because I couldn't answer them at all.*

The student rejected one of those questions herself a minute later, seeing that it did not match the syllabus. The generated practice questions were no better: either irrelevant to the current syllabus or trivial rephrasings of the problem sets, and the exam itself did not resemble anything Claude had produced. The student was already scared of sitting it with nothing behind her but the problem sets, and the irrelevant questions added a second fear, of material that was not on the paper.

With Claude unable to help on any of the three, the student fell back on traditional study: repeating the problem sets, and writing flashcards with definitions, proof steps, and her own step-by-step algorithms for each problem type. *"I just went through the problem sets like 500 times, did them all over again, wrote flashcards, just the typical studying."* Claude still seemed useful for concept explanation during this period, and the chatlogs show the student continuing to use it through the revision weeks.

The revision period brought together several patterns from the rest of the term: the trust and verification cycle from Self-correction, trust, and the limits of verification, when the student caught errors in exam material she could check against her own knowledge (c157, c175), the Socratic friction from How Socratic questioning worked against the student, where the student repeatedly shut down guiding questions in favour of direct answers (c133, c139, c165), and the tension between independent work and Claude use that runs through Doing the mathematics vs asking for it.

4 Discussion

We followed Lena's interactions with her Socratic Claude configuration as she studied for her Topics in Algorithms course across a full teaching term. The chatlogs and the interview, taken together, show a student who used the chatbot productively for much of the course, asking it to explain proof steps, checking its derivations, and correcting its mistakes, although her later prompts showed progressively less of her own prior reasoning. We also observed how the Socratic configuration failed to calibrate its questions to what the student already knew in many exchanges, pitching below her level and above it in the same session. In light of our findings, we return to the three research questions below.

RQ1. How did the student's study practices develop?

In 80 of the 196 exchanges, Lena asked Claude to explain a concept or a step in a proof. In the early sessions she brought her own partial derivations for Claude to check. Over the term, however, her prompts showed less and less of her own prior reasoning. Whereas in January she would work through part of a derivation before asking about one step, by March she would paste a slide title and ask Claude to explain the whole algorithm. Lena herself noticed her change in attitude: *"My frustration level has sunk, like the accepted time I'll think about something before I ask Claude, like, how do I start thinking about this?"*

Lena's reported lower threshold for turning to Claude is consistent with Fan et al. (2025)'s concept of "metacognitive laziness": becoming quicker to seek AI help rather than continuing to work through a difficulty independently. At the same time, offloading did not replace mathematical reasoning: even in later weeks Lena detected missing steps, traced algebra, and checked derivations. What she delegated varied with the mathematical activity, consistent with Yan and Gašević (2026)'s "dynamic agency allocation." By the revision period, Lena had returned to problem sets, flashcards, and working through proofs on her own.

RQ2. When did the chatbot's Socratic configuration support the student's mathematical work, and when did it obstruct it?

Lena often identified the precise step of a proof where she was stuck, and at times the chatbot helped her work through it. The difficulty was not Socratic questioning itself but its fit with the mathematical activity at hand. Guiding questions could support reconstruction, but were obstructive when she wanted to check completed work or resolve one specific step. Effective scaffolding therefore requires knowing both what the student already knows and what mathematical work she is trying to do. These findings extend concerns that prompt-level Socratic guardrails cannot by themselves track a learner's developing understanding (Khosravi et al., 2026; Schell et al., 2025): in this case, effective support also depended on the mathematical activity the student was trying to undertake.

RQ3. How does the student interpret her experience?

For most of the term, Lena valued Claude as less intimidating than contacting teaching staff and better matched to her level than a lecturer would be, even though the chatlogs show the chatbot misjudging that level repeatedly. She reported becoming quicker to turn to Claude rather than continuing to work through a difficulty herself, something she regarded as harmful to her learning. The chatlogs cannot establish how much thinking occurred before each prompt, but later prompts showed less of her own prior reasoning. Her account is consistent with other studies in which participants, students and professionals alike, spent less time reasoning independently when they had access to AI (Fan et al., 2025; Shaw and Nave, 2026).

Her interpretation of Claude's self-corrections also shows the difficulty of mathematical verification. She treated mid-response rewrites as evidence that Claude had checked its working, although no such verification had occurred. When she corrected Claude, by contrast, she had independently checked the lecture notes or carried out calculations herself. In these episodes, using the chatbot critically depended on independent mathematical work rather than on the apparent plausibility of its answer (Zhuang, 2026).

Limitations

The findings above add nuance to patterns reported in the literature, but naturally they are bound to one student's use of one chatbot across a single course. Participation was also self-selected, so the case may not capture students who would be less willing to share their GenAI interactions. The longitudinal record is longer than most comparable studies, covering eleven teaching weeks and a revision period, yet what we describe is specific to this participant's dispositions and learning preferences. Students with different dispositions toward GenAI produce qualitatively different interactions, as other studies have documented (Yoon et al., 2024; Zhuang, 2025; Cardoso-Silva et al., 2025), and a cohort study would be better placed to test whether the longitudinal pattern described here is common or specific to this student's circumstances.

We did not measure learning outcomes. Nor can the chatlogs reveal how much independent thinking occurred before a message was sent. The decline in visible prior reasoning across the term could also reflect the student's growing fluency with the tool, the increasing difficulty of the course material, or time pressures from other courses.

The interview was conducted after the examination period, and what the student told us about her studying may reflect how she thinks about it now as much as what she experienced at the time. We checked the student's claims against the chatlogs where possible, but some are not verifiable from the record.

Our findings concern one implementation of Socratic support, and other models, other configurations of Claude, and differently designed Socratic behaviours could produce different results (Zhuang, 2026).

5 Conclusion

This case study documents one Master's student's use of a Socratic chatbot across a full term of a proof-oriented postgraduate mathematics module at a research-intensive UK university. The longitudinal record, 196 exchanges across eleven teaching weeks and a revision period for the exam, adds temporal detail to patterns that existing studies have documented in isolated tasks, single interventions, or classroom observations. It allows us to trace how the student's use of the chatbot changed as the course progressed and to compare that record with her retrospective account.

Our participant (pseudonym Lena) made Claude part of how she studied across the term. Being new to proof-based mathematics, she worked through the weekly problem sets on her own and often asked the chatbot about particular steps where she got stuck, sometimes because the course assumed prior knowledge she did not have. She brought those questions to the chatbot instead of to a lecturer because they felt too basic, and for most of the term she perceived the GenAI tool as better matched to her level than a lecturer would be, even though the chatlogs show the chatbot misjudging that level repeatedly. As the term went on, her prompts showed less of her own prior mathematical reasoning, while she herself reported becoming quicker to turn to Claude. At the same time, she continued in some exchanges to challenge the chatbot and correct its mathematics using calculations and derivations of her own.

Close reading of the chatlogs also showed that the same form of support was not appropriate for every mathematical activity. Socratic questioning could help when Lena was working towards an argument, but obstructed her when she wanted to check completed work or resolve a specific step. Effective support therefore needs to respond both to what the student already knows and to the mathematical work she is trying to do.

Future work could follow a full cohort longitudinally to test whether these patterns generalise, and test whether support calibrated to both student knowledge and mathematical activity improves mathematical learning.

References

Abdelghani, R., Kaiser, P., and Murayama, K. From Prompting to Epistemic Proactivity: Temporal Trajectories of Student-AI Interaction in Mathematics Learning, June 2026.

Aguilar, M. S. and Puga, D. S. E. Mathematical help-seeking: Observing how undergraduate students use the Internet to cope with a mathematical task. *ZDM*, 52(5):1003–1016, Jan. 2020. ISSN 1863-9690. doi: 10.1007/s11858-019-01120-1.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26): e2422633122, July 2025. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2422633122.

Cardoso-Silva, J., Sallai, D., Kearney, C., Panero, F., and Barreto, M. E. Mapping Student-GenAI Interactions onto Experiential Learning: The GENIAL Framework. *SSRN Electronic Journal*, page 22, Oct. 2025. doi: 10.2139/ssrn.5674422.

Denzin, N. K. and Lincoln, Y. S., editors. *The SAGE Handbook of Qualitative Research*. SAGE, Los Angeles London New Delhi Singapore Washington DC Melbourne, fifth edition edition, 2018. ISBN 978-1-4833-4980-0.

Dilling, F. and Herrmann, M. Using large language models to support pre-service teachers mathematical reasoning—an exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence*, 7, Oct. 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1460337.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., and Gašević, D. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2):489–530, Mar. 2025. ISSN 0007-1013, 1467-8535. doi: 10.1111/bjet.13544.

Flyvbjerg, B. Five Misunderstandings About Case-Study Research. *Qualitative Inquiry*, 12(2):219–245, Apr. 2006. ISSN 1077-8004. doi: 10.1177/1077800405284363.

Iannone, P., Czichowsky, C., and Ruf, J. The impact of high stakes oral performance assessment on students' approaches to learning: A case study. *Educational Studies in Mathematics*, 103(3):313–337, Mar. 2020. ISSN 0013-1954, 1573-0816. doi: 10.1007/s10649-020-09937-4.

Ibrahim, E. I. and Voyer, A. Qualitative research with LLM chatbots: Technological reflexivity for interpretative technology. *Qualitative Research*, 26(1):133–159, Feb. 2026. ISSN 1468-7941, 1741-3109. doi: 10.1177/14687941251390794.

Jin, Y., Yang, K., Martinez-Maldonado, R., Gasevic, D., and Yan, L. The agency gap in AI-supported writing: How reactive and proactive agent designs shape multimodal reasoning. *Computers and Education: Artificial Intelligence*, 11:100655, Dec. 2026. ISSN 2666-920X. doi: 10.1016/j.caeai.2026.100655.

Kempny, C., Frings, J., Rust, P., Meister, S., and Fehring, L. The use and methodological reporting of large language models in qualitative research: A scoping review. *BMC Medical Research Methodology*, 26(1):137, June 2026. ISSN 1471-2288. doi: 10.1186/s12874-026-02913-1.

Khosravi, H., Gasevic, D., Sadiq, S., Yan, L., Lodge, J., Tangen, J., Denny, P., DiCerbo, K., Shum, S. B., and Baker, R. S. Building AI Companions that Prioritise Learning over Performance, 2026.

Klawa, H., Rajpal, S., and Thomas, C. Gen AI in Proof-based Math Courses: A Pilot Study, Sept. 2025.

Locke, K., Kontorovich, I., and Darragh, L. Transforming mathematical identity: Changes in one international student's positioning during first-year mathematics tutorials. *International Journal of Mathematical Education in Science and Technology*, 54(9):1785–1803, Oct. 2023. ISSN 0020-739X. doi: 10.1080/0020739X.2023.2259917.

López-Pernas, S., Misiejuk, K., Oliveira, E., and Saqr, M. The dynamics of the self-regulation process in student-AI interactions: The case of problem-solving in programming education. In *Proceedings of the 25th Koli Calling International Conference on Computing Education Research*, pages 1–12, Koli Finland, Nov. 2025. ACM. ISBN 979-8-4007-1599-0. doi: 10.1145/3769994.3770043.

Makransky, G., Shiwalia, B. M., Herlau, T., and Blurton, S. Beyond the "Wow" Factor: Using Generative AI for Increasing Generative Sense-Making. *Educational Psychology Review*, 37(3):60, Sept. 2025. ISSN 1040-726X, 1573-336X. doi: 10.1007/s10648-025-10039-x.

McGrath, C., Palmgren, P. J., and Liljedahl, M. Twelve tips for conducting qualitative research interviews. *Medical Teacher*, 41(9):1002–1006, Sept. 2019. ISSN 0142-159X, 1466-187X. doi: 10.1080/0142159X.2018.1497149.

Misiejuk, K., López-Pernas, S., Kaliisa, R., and Saqr, M. Cognitive offloading in student–AI collaboration: A longitudinal analysis of prompting strategies. *Computers in Human Behavior Reports*, 22:101130, May 2026. ISSN 2451-9588. doi: 10.1016/j.chbr.2026.101130.

Neuhaus-Eckhardt, S., Sommerhoff, D., Rach, S., and Ufer, S. Proof Comprehension, Proof Validation, & Proof Construction: All the same or Different Skills? An Answer Based on an Empirical Analysis. *International Journal of Science and Mathematics Education*, 23(8):4153–4178, Nov. 2025. ISSN 1571-0068. doi: 10.1007/s10763-025-10621-3.

Risko, E. F. and Gilbert, S. J. Cognitive Offloading. *Trends in Cognitive Sciences*, 20(9):676–688, Sept. 2016. ISSN 13646613. doi: 10.1016/j.tics.2016.07.002.

Rismanchian, S., Uzun, H., Matayoshi, J., Cosyn, E., and Kurd-Misto, E. Faster Completion, Less Learning: Generative AI Reduced Study Time on Math Problems and the Knowledge They Build, July 2026.

Schell, J., Ford, K., and Markman, A. B. Building responsible AI chatbot platforms in higher education: An evidence-based framework from design to implementation. *Frontiers in Education*, 10:1604934, July 2025. ISSN 2504-284X. doi: 10.3389/feduc.2025.1604934.

Shaw, S. D. and Nave, G. Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender, Jan. 2026.

Urhan, S. and Zhuang, Y. Can ChatGPT Support Proof Validation? Exploration through Habermas' Construct of Rationality. *International Journal of Research in Undergraduate Mathematics Education*, Apr. 2026. ISSN 2198-9745. doi: 10.1007/s40753-026-00300-1.

Yan, L. and Gašević, D. Agentivism: A learning theory for the age of artificial intelligence, Apr. 2026.

Yoon, H., Hwang, J., Lee, K., Roh, K. H., and Kwon, O. N. Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education*, 56(7):1531–1551, Aug. 2024. ISSN 1863-9690. doi: 10.1007/s11858-024-01629-0.

Yoon, H., Lee, Y., Lee, K., and Kwon, O. N. Before or after? Beyond timing of students' interaction with generative artificial intelligence in proving mathematical statements. *ZDM – Mathematics Education*, May 2026. ISSN 1863-9690. doi: 10.1007/s11858-026-01800-9.

Zhuang, Y. Lessons from Using ChatGPT in Calculus: Insights from Two Contrasting Cases. *Journal of Formative Design in Learning*, 9(1):25–35, Mar. 2025. ISSN 2509-8039. doi: 10.1007/s41686-025-00098-2.

Zhuang, Y. Empowering students to critically validate AI-generated mathematical solutions through the rational questioning approach. *Educational Studies in Mathematics*, June 2026. ISSN 0013-1954. doi: 10.1007/s10649-026-10528-y.

Zhuang, Y. and Conner, A. Teachers' use of rational questioning strategies to promote student participation in collective argumentation. *Educational Studies in Mathematics*, 111(2):345–365, July 2022. ISSN 0013-1954. doi: 10.1007/s10649-022-10160-6.