International Journal of Artificial Intelligence in Education (2025) 35:3896–3921
https://doi.org/10.1007/s40593-025-00521-6

ARTICLE

You've Got AI Friend in Me: LLMs as Collaborative Learning Partners

Daniel M. Oppenheimer[1,2] · **Trent N. Cash**[1,2] · **Allison E. Connell Pensky**[3]

Received: 5 September 2024 / Accepted: 21 September 2025 / Published online: 10 November 2025
© The Author(s) 2025

Abstract

Collaborative learning - an educational paradigm in which students work together to master content and complete projects - has long been a staple of classroom pedagogy. The rapid development of Large Language Model chatbots (LLMs), such as ChatGPT, Gemini, and Claude, has created the potential for a new frontier in collaborative learning in which students collaborate not with other students, but with conversational LLMs. In this exploratory, classroom-based study, we tested a novel intervention in which college students taking an introductory level social science elective ($n = 154$) were asked multiple times throughout a semester to write argumentative essays, have them critiqued by LLMs, and improve their essays by incorporating the LLMs' critiques into their arguments or rebutting the LLMs' arguments. We found that students engaged deeply with the LLMs, enjoyed working with LLMs, and that their argumentative writing skills improved over the course of the semester. This improvement was noted even in the essays that students wrote without the support of LLMs. Throughout the semester, students also improved in their prompt engineering skills and showed increased self-efficacy for working with LLMs, however future research is needed to isolate the causal effects of the intervention. Implications and limitations of the intervention are discussed.

Keywords Collaborative learning · Large language models · Argumentation · Argumentative writing · Student-AI collaboration

✉ Trent N. Cash
trentncash@gmail.com

Allison E. Connell Pensky
aconnel2@andrew.cmu.edu

[1] Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[2] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

[3] Eberly Center for Teaching Excellence & Educational Innovation, Carnegie Mellon University, Pittsburgh, PA, USA

Introduction

Education is often thought of as the transmission of knowledge from an instructor to students. However, educational research from recent decades has highlighted the numerous benefits of collaborative learning, an educational approach in which students are encouraged to work together to master content, solve problems, and complete projects (Hmelo-Silver & Chinn, 2016; Laal & Laal, 2012; Nokes-Malach et al., 2015). Historically, collaborative learning has been thought of as a collaboration between students – typically in a classroom, with students working in pairs or small groups. However, collaborative learning need not be exclusively between humans; in fact, much of modern society requires collaboration between humans and technology (Baroroh et al., 2021; Sowa et al., 2021) – and learning is no exception (Alam, 2022; Kim et al., 2022; Su et al., 2023).

Today, one of the most important technological tools that students can collaborate with is artificial intelligence (AI). Due to their cultural salience, the AIs that students are most likely to encounter are high-profile, generative Large Language Models (LLMs), such as ChatGPT (OpenAI, 2024) and Gemini (Google, 2024). LLMs are AI models that are designed to understand and respond to human language - an ability known as natural language processing (NLP; Chowdhary, 2020; Nadkarni et al., 2011). This is achieved by training the model on a large corpus of text so that it can learn how language is structured and what words frequently co-occur. Modern LLMs are trained on large swaths of the internet and contain billions - and in some cases trillions - of parameters that make them highly effective at interpreting and generating human language (OpenAI, 2024; Roose, 2023; Wu et al., 2023). Based on this knowledge, LLMs can accept prompts and use their parameters to determine what a good (i.e., well-reasoned, accurate) response to the prompts would be (Roose, 2023; Wolfram, 2023). For a dedicated review of the history and use of LLMs, see Wu et al. (2023).

These LLMs are distinct from the educational technology of the past because they are adaptive, maintain knowledge of a wide variety of domains, and, perhaps most importantly, respond to prompts in a conversational tone that mimics that of human collaborators. For example, LLMs can take on personas (Desai et al., 2025; Hu & Collier, 2024) adjust their responses to match the perspectives of users (Sun & Wang, 2025), adhere to social norms (Horiguchi et al., 2024), and even express some forms of empathy (Sorin et al., 2024). Indeed, LLMs can mimic human behavior to such an accurate extent that users struggle to identify whether content is generated by a human or LLM (Radivojevic et al., 2024). Given these human-like characteristics, students may be able to collaborate with LLMs in similar ways as they collaborate with human peers. LLMs' conversational nature also makes them distinct from existing AI-based educational technologies that prioritize instruction through adaptive feedback and tutoring (e.g., Alam, 2022; Vidergor & Ben-Amram, 2020). Furthermore, LLMs offer highly accessible user interfaces that make them easy and fun to use, both in the classroom and in students' free time (Cash & Oppenheimer, 2024; Su et al., 2023; Zhang & Tur, 2024).

While research has been rapidly emerging on the ways in which students use LLMs and the potential impacts that they have on attitudes and learning, there is still

a stark lack of consensus in the literature. Different scholars interpret and describe students' use of LLMs in different ways. A recent meta-analysis (Wang & Fan, 2025) categorized the educational use cases for LLMs into three distinct roles: personalized tutors (e.g., giving feedback as an instructor would), intelligent partners (e.g., acting more like a peer or classmate), or intelligent learning tools (e.g., serving as a glorified search engine). Here, we expand on the idea of interacting with LLMs as intelligent partners by embedding the core features of collaborative learning into student-LLM interactions and utilizing LLMs' conversational strengths to enhance students' learning experiences. Further developing empirical evidence about how students interact with LLMs is critical to designing effective LLM-based educational interventions.

In the study presented here, students collaborated with LLMs throughout the semester by asking them for feedback, requesting revisions, and having iterative conversations as they developed argumentative essays in which they were tasked with articulating positions on course-relevant issues. At the beginning and end of the semester, we measured students' self-reported, open-ended reactions to the assignment (e.g., how the assignment was useful, what parts of the assignment were challenging), self-efficacy for using AI (e.g., their confidence in their ability to generate effective prompts and evaluate LLM output), affective and motivational responses to the assignment (e.g., how useful, challenging, engaging, and enjoyable the assignment was), and actual performance outcomes, including the quality of students' arguments, rebuttals, and prompt engineering (i.e., their usage of strategies to maximize the quality of LLM responses). To foreshadow results, students engaging in these scaffolded assignments had overwhelmingly positive affective and metacognitive reactions to the assignments, and, while we cannot confidently isolate the specific causal pathway, showed learning gains over the course of the semester.

Collaborative Learning

The ability to collaborate is one of the great strengths of humankind. By pooling resources, knowledge, and effort, groups of individuals can often produce outcomes that are more than the sum of their parts (Melis, 2013). The benefits of collaboration have been highlighted in numerous disparate domains, such as policymaking (Connick & Innes, 2003), business (Hoegl & Gemuenden, 2001), academia (Lewis et al., 2012), and the creative arts (Sawyer & DeZutter, 2009). It should be unsurprising, then, that collaboration has been found to be an effective tool for improving educational outcomes (Andrews & Rapp, 2015; Nokes-Malach et al., 2015; Rajaram & Pereira-Pasarin, 2010). Whether or not collaboration with LLMs similarly positively impacts educational outcomes is an open question.

When collaboration is discussed in educational settings, it is typically referred to as collaborative learning. While there is no established consensus definition of collaborative learning, the existing definitions generally center around three key components: (1) The learner must be working with an external entity; (2) The learner and external entity must have a shared goal or outcome; and (3) The goal or outcome must be educational in nature (Chinn & Clark, 2013; Hmelo-Silver & Chinn, 2016; Laal & Laal, 2012; Nokes-Malach et al., 2015). In its most common form, collab-

orative learning occurs when students work in pairs or groups (Component #1), to study a complex topic or complete a group project (Component #2), with the goal of mastering the classroom materials (Component #3). Collaborative learning has been demonstrated to have wide-ranging benefits, including improvements in learning outcomes (Kirschner et al., 2011; Terenzini et al., 2001; Van Boxtel et al., 2000), classroom engagement (Sun et al., 2022), perspective taking (Kim et al., 2024), argumentative reasoning (Kuhn & Crowell, 2011), problem solving (Lazakidou & Retalis, 2010), and peer communication (Lin et al., 2022; Terenzini et al., 2001).

Researchers have identified multiple mechanisms through which collaborative learning generates these benefits. One major mechanism is elaboration, the process by which a student verbally shares their knowledge with their collaborators (Hmelo-Silver & Barrows, 2008; Van Boxtel et al., 2000). The most obvious benefit of elaboration is that it allows students to co-create knowledge – generating opportunities for students to learn from what other students know. The elaborating student also benefits from talking through their ideas because doing so can help them fill in gaps in their reasoning and develop deeper knowledge of how concepts are related (Andrews & Rapp, 2015; Van Blankenstein et al., 2013; Webb, 1991). Elaboration also benefits the speaker because it forces students to engage with knowledge repeatedly, thus reinforcing their memory of what was learned (Congleton & Rajaram, 2011; Rajaram & Pereira-Pasarin, 2010). As such, elaboration encourages deeper thinking about a topic for both speakers and their collaborators. Because of their conversational nature, LLMs are well-equipped to engage in iterative bi-directional elaboration. Students elaborate their thoughts and beliefs through the prompts they provide, then the LLMs elaborate through the responses they generate to those prompts. Students can also specifically request that LLMs elaborate about a specific topic that they wish to learn more about.

A second major mechanism of collaborative learning is argumentation, the process by which students debate the merit of each other's ideas and resolve disagreements or confusion about how the knowledge is (or should be) structured (Chinn & Clark, 2013; D'Mello et al., 2014; Nussbaum, 2008). The argumentation process allows learners to fill gaps in their knowledge with the knowledge of their collaborators (Larrain et al., 2021; Retnowati et al., 2018) and to correct any misunderstandings they have about the content or the task (Chin & Osborne, 2010; Gadgil & Nokes-Malach, 2012). Argumentation can also help students to develop deeper knowledge of their own beliefs through the process of trying to justify their beliefs to their collaborators who have different perspectives than their own (Chin & Osborne, 2010; Kim et al., 2024; Nussbaum, 2008). Importantly, argumentation is only effective when a student works with collaborators that do not simply agree with and support them, but rather challenge them to think harder and consider alternatives (Chinn & Clark, 2013; Gadgil & Nokes-Malach, 2012; Kim et al., 2024). In the present study, students were tasked with prompting LLMs to engage in this kind of collaborative learning through argumentation by asking for critiques of their human-generated work.

Challenges of Collaborative Learning

Despite its numerous benefits, collaborative learning also presents unique challenges. The most obvious challenge is that collaborative learning groups may not get along. Reasonable disagreements may turn into combative arguments, lower-status or lower-ability students may be excluded from the group, or individual students may become domineering (Cohen et al., 1999; Lee et al., 2015; Näykki et al., 2014). If groups do not coordinate effectively, they are less likely to benefit from collaboration. A second major challenge is the tendency for individual group members to disengage from the collaborative learning process, forcing collaborators to shoulder the bulk of the burden – a process known as social loafing (Gabelica et al., 2022; North et al., 2000). When students socially loaf, they do not engage in the elaborative and argumentative processes necessary for collaborative learning to succeed, hampering both their own learning and the learning of their collaborators (Chin & Osborne, 2010; Gadgil & Nokes-Malach, 2012; Larrain et al., 2021; Retnowati et al., 2018). These interpersonal challenges are not unique to the educational domain, but rather manifestations of human social processes that appear in nearly all domains. However, similar issues are unlikely to arise when collaborating with a technological partner (e.g., LLMs), potentially allowing students to reap the benefits of collaborative learning while escaping the detriments of interpersonal challenges.

Technology as a Collaborator

Given the benefits and challenges inherent in collaborating with other humans, educators have long been interested in understanding how students may be able to collaborate with technology instead (Kim et al., 2022; Koedinger et al., 1997; Lipponen & Lallimo, 2004; Su et al., 2023). In recent decades, there has been growing interest in adaptive tools that tailor learning to students' needs using complex algorithms or artificial intelligence, such as those used by tutoring organizations like Khan Academy (Vidergor & Ben-Amram, 2020). These adaptive tools have been found to be highly effective at promoting content mastery and remediating students who are falling behind their classmates (Alam, 2022; Koedinger et al., 1997; Latham et al., 2014), but they are typically designed to be instructional programs, not dialogic partners with whom students can engage in true collaboration. As a result, students are likely to perceive these tools as experts or teachers who are transmitting knowledge, not as collaborators.

True dialogic collaboration with educational technology only recently became possible with the mass popularization of powerful generative AI chatbots built on Large Language Models (LLMs), such as ChatGPT and Gemini. These LLMs differ from other AI-based educational technologies in that they are designed as general purpose conversational models, not strictly pedagogical tools. Thus, LLMs engage with users in ways that mimic normal human interaction. For example, they default to a friendly tone (Sun & Wang, 2025); engage in pleasantries (Zhao & Hawkins, 2025); challenge users' misconceptions (Costello et al., 2024); ask users for clarification when a prompt is ambiguous (Zhang & Choi, 2023); remember previous

prompts within conversations (Wang et al., 2025a, b); provide meaningful confidence judgments about the accuracy of their claims (Cash et al., 2025); adhere to social norms (Horiguchi et al., 2024); demonstrate creativity (Zhao et al., 2025); and communicate using language that is difficult to distinguish from human-generated text (Radivojevic et al., 2024). As a result, working with an LLM may feel less like being tutored or tested, and more like chatting with another human. This key distinction could allow students to interact with LLMs in ways that mimic collaborative learning with peers. Of course, one notable difference between collaborative learning with humans and LLMs is that human-human collaboration is typically reciprocal (i.e., both students learn; Nokes-Malach et al., 2015), whereas human-LLM interaction is not inherently beneficial to the LLM (aside from data acquisition).

In the educational context, many scholars have treated LLMs as a nuisance because of their potential to facilitate large-scale cheating and plagiarism (Milano et al., 2023). However, LLMs also provide an excellent opportunity for students to engage in collaborative learning without the need for other students (Kim et al., 2022; Su et al., 2023). For example, students could ask LLMs to review their work, provide feedback on their understanding of a topic, or give suggestions on elements of a project that the students themselves lack the skills to complete – all of which emulate ways in which they might work with a human collaborator. Because of their conversational nature, LLMs can engage in these collaborative tasks in a way that may feel more natural to students. However, the empirical literature evaluating the efficacy of collaboration between students and LLMs is recent and nascent, and the questions are plentiful. One key question is how effectively students can collaborate with LLMs in ways that promote elaboration and argumentation.

There are reasons to believe that collaboration with LLMs can facilitate student learning. At the most basic level, LLMs provide an engaged partner with whom students can share their thoughts, allowing the student to take on the role of elaborator and think through their understanding of the material (Andrews & Rapp, 2015; Van Blankenstein et al., 2013; Webb, 1991). Moreover, LLMs do not – or at least, should not – engage in the problematic social behaviors, such as social loafing (Gabelica et al., 2022; North et al., 2000), combativeness, and social exclusion (Cohen et al., 1999; Lee et al., 2015; Näykki et al., 2014), that hinder collaborative learning among humans.

On the other hand, one significant concern about LLMs arises when they take on the role of elaborator. LLMs have a propensity to provide inaccurate or made-up information (Chen et al., 2023), which could confuse or misinform students (LLM hallucinations have also received extensive coverage in the news media, e.g., Weise & Metz, 2023). For this reason, successful collaboration with LLMs requires that students maintain a healthy dose of skepticism that their technological collaborator may be misinformed - just like a human collaborator may be – reinforcing the need for maintaining their own critical thinking. Furthermore, LLMs may be unwilling to contradict students' misunderstandings or disagree with students about their beliefs – both of which are critical components of collaborative learning and argumentation (Chin & Osborne, 2010; Gadgil & Nokes-Malach, 2012; Kim et al., 2024; Nussbaum, 2008).

Moreover, there is ample evidence that students who offload work to assistive technology - rather than engaging in the cognitive effort to complete tasks themselves – show learning deficits (Hamilton & Benjamin, 2019; Hargis, 2019). If students use LLMs to avoid engaging with the task and processing the content, such as by having an LLM write a paper for them, then the collaboration may negatively influence learning outcomes. Similarly, if students treat LLMs as authoritative experts and blindly accept whatever the LLM says, then they are unlikely to reap the benefits of collaborative learning, which relies on bi-directional communication (Andrews & Rapp, 2015; Hmelo-Silver & Barrows, 2008; Nokes-Malach et al., 2015; Van Boxtel et al., 2000). For this reason, LLMs are likely to function most optimally as collaborative partners if students view them as peers with whom they can exchange knowledge, rather than experts who are transmitting knowledge (Erkens & Bodemer, 2019; Van Boxtel et al., 2000). Thus, collaborative learning interventions using LLMs may benefit from encouraging students to approach LLMs not as experts or reference tools, but rather as sounding boards for thoughtful discussion and deliberation.

Students may also simply dislike working with LLMs. For example, one study suggests that participants rate collaborators who are known to be AI as less creative, less intelligent, and less likable than AI collaborators that participants are told are humans (Ashktorab et al., 2020). However, emerging research on the use of generative AI in higher education suggests that students hold generally positive attitudes about working with an LLM in their classes and can become critical consumers of its outputs. For instance, during a focus group, Essel et al. (2024) found that students felt positively about ChatGPT (e.g., it increased their confidence; it used human-like interactions; it was efficient) but they were also able to critically highlight its weaknesses (e.g., providing inaccurate information and citations). Mahapatra (2024) also reported from a focus group that students overwhelmingly found ChatGPT to be a positive support for their writing (e.g., content, organization, and grammar) while recognizing its constraints (e.g., formulaic writing patterns).

Despite the many potential pitfalls highlighted here, if students elaborate their beliefs to LLMs, work through disagreements with LLMs, and ask LLMs to fill gaps in their knowledge, they may benefit similarly to how they would if they were working with a human collaborative learning partner. Plausibly, in order to maximize the benefits of collaboration with LLMs, students would need to be trained in how to properly collaborate with an LLM – just like they need to be taught how to collaborate with their human peers (Salmons, 2023). This could be achieved through direct instruction (e.g., training students on how to effectively generate prompts) or scaffolding (e.g., structuring assignments to force students toward effective use of LLMs, and having students practice this skill by completing the assignment multiple times throughout a semester), depending on the needs of the students.

Argumentative Reasoning

While there are many educational domains in which LLMs may be effectively used as collaborative learning partners, one that may be of particular interest is argumentative reasoning (Su et al., 2023), which can be broadly defined as “the construction

and evaluation of arguments" (Iordanou et al., 2016, p. 39). In educational contexts, argumentative reasoning is often taught through written assignments that allow students to practice creating cohesive, evidence-based arguments (Ferretti & Graham, 2019; Preiss et al., 2013). Toulmin's (2003) theory of argumentation highlights six components of an argument that a student should learn to generate: (1) The **claim** the student wants to make; (2) The **grounds** (i.e., evidence) that support the claim; (3) The **warrant**, which explains the connection between the claim and the grounds; (4) The **backing**, which further justifies the warrant (e.g., an example of the warrant in practice); (5) The **qualifier**, which highlights circumstances under which the claim may or may not be true; and (6) The **rebuttal**, which anticipates and/or responds to counterarguments.

Decades of research suggest that collaborative learning can be a useful tool for promoting the development of the (meta-)cognitive skills necessary to engage with each of these components of argumentative reasoning (Nussbaum, 2008; Sampson & Clark, 2009). When students elaborate their beliefs to their collaborators, they learn to make specific claims, provide grounds for their claims, and explain how their evidence warrants their claims in a way that their collaborators will understand and believe (Chinn & Clark, 2013; Chinn et al., 2000; Jonassen & Kim, 2010; Kim et al., 2024; Lin et al., 2019; Maloney & Simon, 2006). Then, when students engage in the argumentation component of collaborative learning, they are forced to consider the perspectives of their collaborators and identify gaps in their reasoning in order to revise their arguments and/or generate effective rebuttals (Chin & Osborne, 2010; Felton et al., 2015; Jonassen & Kim, 2010; Kim et al., 2024; Lin et al., 2019; Osborne, 2010; Wen et al., 2023). Students may also be nudged to accept qualifiers to their claim that they would not have considered on their own (Chin & Osborne, 2010; Kuhn, 1991; Weinberger & Fischer, 2006). Collaborative learning may be particularly adept at promoting argumentative reasoning because, some scholars argue, the core purpose of argumentation is to exchange arguments with others (Mercier, 2016).

The development of argumentative reasoning skills can be enhanced through the use of educational technology (Lee et al., 2024). Many of these technological advances have focused on promoting collaborative learning by digitally connecting students through tools such as discussion boards (Scheuer et al., 2010; Weinberger & Fischer, 2006). Advanced technologies can also be used as argumentative writing tutors that aid students in crafting arguments (Hoffmann, 2015) and provide adaptive feedback on the quality of arguments generated by students (Scheuer et al., 2010; Wambsganss et al., 2020). More recently, the literature has shifted its focus to LLMs. Su et al. (2023) argue that LLMs can serve multiple roles in promoting argumentative reasoning skills, such as providing evaluative feedback or engaging as a collaborative peer. Though the empirical evidence is nascent, early research suggests that LLMs can provide meaningful feedback on undergraduate-level argumentative writing (Wang et al., 2024), effectively assess the quality of arguments (Mirzakhmedova et al., 2024), re-write argumentative claims to make them less ambiguous (Y. Wang et al., 2025a, b), and correct erroneous or fallacious beliefs held by human reasoners (Costello et al., 2024). Given these skills, LLMs present a promising avenue for enhancing students' argumentative reasoning skills through collaborative learning interactions.

Student Use of Generative AI in the Writing Process

In the study presented here, we focus specifically on how students collaborate with LLMs when engaging in argumentative reasoning in a written form. Recent research on students' use of generative AI tools in the writing process has explored the ways that students choose to use the tools, their attitudes about the tools, and the potential impact that using LLMs has on them. For instance, Usdan et al. (2024) found that graduate students could prepare a policy memo significantly faster with the help of LLMs. This type of 'productivity effect' has been observed by other researchers as well. Noy and Zhang (2023) similarly found that producing a written deliverable was accomplished significantly faster with genAI, especially for lower-performing individuals. However, they found that these participants were largely copying and pasting the genAI output rather than cognitively engaging with the task. These studies did not include a transfer task in which the writing process and deliverable were completed without generative AI in order to test if the participants experienced a skill boost along with the time savings.

Other researchers have focused on the quality of the written deliverable rather than the speed with which it was produced. For instance, Mahapatra (2024) tested whether English-as-a-second-language students' use of ChatGPT as a writing feedback tool would impact the quality of their English writing deliverables. They found that using ChatGPT improved students' pre-to-post writing and that students' attitudes about the tool were favorable in terms of supporting their writing content, organization, and grammar. Fan et al. (2024) evaluated the benefits of using genAI to assist in revising written essays. They found that using LLMs led to greater improvements in essay quality than feedback from a human expert, use of a checklist, or no intervention at all. However, by investigating learning trace data they found that the LLM group was overly reliant on the tool and exhibited less connection to their revisions process than the control or human expert groups. This result prompted the authors to caution against potential tool dependency and "metacognitive laziness" when using LLMs. These initial forays into the benefits and risks of using LLMs in writing have provided useful scaffolding for follow up research, but given the nascence of the domain, there remain many unexplored questions about the use of LLMs in student writing and the associated cognitive processes.

Current Study

To explore the extent to which LLMs could effectively function in the role of collaborative learning partners to promote the development of argumentative reasoning skills, we implemented a series of assignments designed to scaffold student collaboration with LLMs, optimizing the features of collaboration (i.e., elaboration and argumentation) that have been shown to be effective for learning in all-human teams, and minimizing the features that are theorized to undermine learning (e.g., social loafing). Five times over the course of one semester, students were assigned to write argumentative essays. Each iteration tasked students with (1) taking and defending a position on an open-ended topic, (2) having an LLM provide critical feedback on the argument

quality, and (3) updating the original argument in response to the feedback of the LLM (either rebutting or incorporating the criticisms to improve the original argument). After the first and fifth iterations of the assignment we measured students' open-ended reactions to the assignment (e.g., how the assignment was useful, what parts of the assignment were challenging), self-efficacy for prompting LLMs and evaluating their output, affective and motivational responses about how useful, challenging, engaging, and enjoyable the assignment was, and actual performance outcomes, including argument quality, rebuttal quality, and prompt engineering quality (i.e., their usage of strategies to elicit helpful LLM responses), as graded by a course teaching assistant. Using these measures, we sought to answer the following research questions:

1. To what extent can LLMs serve as partners for students in collaborative learning activities? That is, can students collaborate with LLMs in ways that have been shown to be effective in collaboration with human peers?
2. Do students who collaborate with LLMs achieve learning gains akin to those found in the collaborative learning literature? More specifically, will students improve their argumentation skills, prompt engineering skills, and ability to respond to feedback in a course that implements scaffolded LLM-based collaborative learning assignments?
3. How will students affectively and metacognitively respond to scaffolded LLM-based collaborative learning assignments in real-world classrooms?

Methods

Participants One hundred and fifty-seven students enrolled in an introductory level social science course focused on the topic of intelligence at Carnegie Mellon University in the fall of 2023 took part in this study as part of their standard coursework. Three students dropped the class during the semester, leaving one hundred and fifty-four participants. During the semester, students were asked to complete two optional surveys to evaluate their experiences with the intervention. While 85 and 100 students completed the first and second survey, respectively, some students left certain questions unanswered, such that the sample size for individual questions ranged from 80 to 97 participants. Participant demographics were not collected at the individual level; however the class was representative of the university as a whole: primarily ages 18–23 years, 57% male, and a racial breakdown of approximately 47% Asian, 42% White, 13% Hispanic, 8% Black, and 6% other (Carnegie Mellon University, 2024; numbers do not add to 100% because some students identify in multiple categories).

Materials and Procedure The primary pedagogical intervention was a series of assignments which provided students with analytic prompts and asked them to write 300-word argumentative essays in which they were told to "formulate [their] own opinion[s], and defend [them] with logic and evidence." Students completed the assignment five times throughout the semester, but we will only evaluate the first and last iteration. The prompt for the first iteration was "What is intelligence?" and the prompt for the final iteration was "Are there questions about intelligence that are inappropriate to study due to the potential social ramifications?" These prompts were

selected because they were germane to the course content of the immediately preceding two weeks. The writing portion of the assignment was designed to mimic the elaboration element of collaborative learning in which a student describes what they know and what they believe to be true (Hmelo-Silver & Barrows, 2008; Van Boxtel et al., 2000) and to provide an opportunity for students to practice making claims, supporting them through grounds and warrants, developing proper qualifiers for when their arguments are valid, and rebutting opposite views (Toulmin, 2003).

Students were then asked to provide their argument to an LLM chatbot of their choice, such as ChatGPT, Gemini, or Claude, and to ask the LLM to critique their argument as thoroughly as possible. Students were informed that they would be evaluated, in part, on the quality of their prompt engineering, assessed via the quality of the rebuttal that they were able to elicit from the LLM. Given that the purpose of this stage of the task was to have the generative AI model challenge the student's beliefs, it can be interpreted as the student engaging in the argumentation element of collaborative learning with the LLM (Chinn & Clark, 2013; D'Mello et al., 2014; Nussbaum, 2008) with a secondary goal of helping students improve at prompt engineering. At the start of the semester, students were provided with a two-page document that provided tips on how to use LLMs effectively. This document, titled "Guidelines for Prompt Engineering," was generated by ChatGPT and edited by the teaching team (for the complete text see Supplementary Materials).

Finally, students were asked to respond to the LLM's argument (including fact checking the accuracy of its position). The wording of the assignment allowed for students to either rebut the arguments made by the LLM by clarifying and expanding upon points of continued disagreement, or to update their original argument if they were persuaded by the LLM's critique. This stage was meant to force students to actively engage with AI feedback and mimicked the iterative nature of the argumentation element of collaborative learning (Chin & Osborne, 2010; Kim et al., 2024; Nussbaum, 2008). For complete wording of the assignment, please see the Supplementary Materials.

At the end of the assignment, students turned in their original argument, a screenshot of their conversation with the LLM, and their updated argument. Assignments were graded by the course teaching assistant, who provided feedback to the students in less than a week. Assignments were graded on a 10-point scale, with up to three points awarded on the basis of how effectively students responded to the initial writing prompt, up to three points awarded on the basis of how effectively they used the LLM (i.e., prompt engineering), up to three points awarded on the basis of the effectiveness of their rebuttal to the LLM or their ability to update their argument based on the LLM's feedback, and up to one point for "professionalism" (i.e., proper grammar, formatting, etc.). As every student received the full point for professionalism on every assignment, it will not be discussed further in this report. In addition to evaluating the quality of the feedback that LLMs provided, the quality of students' prompt engineering was assessed based on their adherence to best practices outlined in the "Guidelines for Prompt Engineering" document provided to them at the beginning of the semester (see Supplemental Materials). These guidelines included concepts such as using clear and precise language, providing the LLM with contextual

information, implementing strategies to avoid misinformation, and testing iterative variants of prompts until the desired result is achieved.

About a week after completing each assignment, students were invited to fill out a brief survey in which they could reflect on their experience. The survey asked several open-ended questions which prompted students to provide feedback about: (1) How the assignment was useful, (2) How the assignment was challenging or frustrating, and (3) Additional comments or experiences.

After each assignment, we measured students' self-efficacy for using LLMs – which we referred to in the assignments as generative AI. As no established self-efficacy for AI scales existed at the time of the study, and one of the most important features of self-efficacy is that it is domain-specific, we created a bespoke self-efficacy scale for the present study, drawing on best practices from the self-efficacy literature (e.g., Bandura 1977; Pajares et al., 2001; Talsma et al., 2018; Tompson & Dass, 2000). At each timepoint, self-efficacy for using generative AI was measured both retrospectively, by asking students to think about their ability to do different tasks prior to completing the assignment, and presently, by asking students to think about their current ability after having completed the assignment. Students used a slider to indicate on a 0–100% scale their confidence in their abilities to (1) "effectively ask a question that would get generative AI (such as ChatGPT or Bard) to answer in the way that [they] wanted", (2) "evaluate the quality of generative AI outputs", and (3) "tell the difference between a prompt that was likely to be effective from a prompt that was likely to be ineffective."

Students were then asked to rate how much experience they had with generative AI prior to the assignment on a −5 (no experience whatsoever) to 5 scale (a great deal of experience). Students then rated how useful, challenging, engaging, and enjoyable the assignment was, again on a −5 to 5 scale. Finally, students were asked how similar the assignment was to previous assignments that they had completed in other classes (−5 (not at all) to 5 (extremely) scale) and how much time the assignment had taken relative to their expectations (−5 (less than I expected) to 5 (more than I expected scale). Bi-valenced Likert scales were used based on a large literature in psychometrics and public opinion research showing that, relative to standard, univalenced scales, they allow for cleaner interpretations of how numeric values map onto constructs that could take on a positive or negative valence (Schwarz, Grayson, & Gnauper, 1998; Schwarz et al., 1991). Students were also provided an open-ended text box in which they could provide any other feedback about the assignment or the survey that they would like.

Students completed the full assignment (independent drafting, engaging with the LLM's position, and revising their position) five times during the course, with their lowest score being dropped. Thus, students could opt out of one iteration of the assignment if they faced time constraints or were happy with their scores on other assignments. Both assignments (Assignment 1 as the pre-test and Assignment 5 as the post-test) evaluated in this study were graded by the same teaching assistant (a human coder) who was blind to the nature of the study but was not blind to the timing of the assignment, nor the specifics of the prompt that students were responding to. The coder evaluated the argumentative writing performance by assessing the Quality of Initial Argument (cf. Münchow, et al., 2019). The coder also evaluated the Quality of Prompt Engineering, and Quality of Response to AI Feedback for each assignment.

Results

Affective Responses After completing the first iteration of the assignment, students self-reported that the assignment was useful ($\bar{x}$ = 2.32), challenging ($\bar{x}$ = 1.42), engaging ($\bar{x}$ = 2.51), and enjoyable ($\bar{x}$ = 2.03). As the reports were on a −5 (not at all) to 5 (extremely) scale, one-sample t-tests comparing these numbers to the neutral midpoint of 0 were run and revealed that they were all highly significant ($ts(79) > 7.84$, $ps < 0.001$). After completing the fifth iteration of the assignment, students self-reported that the assignment was still useful ($\bar{x}$ = 2.48), challenging ($\bar{x}$ = 0.63), engaging ($\bar{x} = 2.22$), and enjoyable ($\bar{x} = 1.81$). The one-sample t-test comparing these numbers to the neutral midpoint of 0 revealed that these were also highly significant ($ts(96) > 3.06$, $ps < 0.002$). None of these metrics significantly changed from the first to fifth iteration of the assignment.

Self-Efficacy for Using Generative AI Only data from students who answered both the retrospective (confidence before the assignment) and current self-efficacy for using generative AI (confidence after the assignment) questions is presented ($n = 81$ for Assignment 1, $n = 98$ for Assignment 5). The three items measuring self-efficacy were combined to form a single inventory dependent measure for each timepoint (retrospective and current) measured for each assignment ($\alpha s = 0.75 - 0.86$). As shown in Table 1, self-efficacy increased across all dimensions between pre-and post-assignment measures. A within-subject t-test comparing the difference between students' retrospective and current self-efficacy revealed that this trend was marginally significant ($t(80) = -1.92$, $p = .06$) for the first iteration of the assignment and highly significant for the fifth iteration of the assignment ($t(97) = -5.65$, $p < .001$). Moreover, students' current self-efficacy was significantly higher after the fifth iteration of the assignment than after the first iteration of the assignment ($t(177) = 4.25$, $p = .006$), suggesting growth throughout the semester.

Table 1 Self-Efficacy descriptive data

Survey Item Scale: 0 (completely uncertain) to 100 (completely certain)	Assign 1: Retro Mean	Assign 1: Current Mean	Assign 5: Retro Mean	Assign 5: Current Mean
I could effectively ask a question that would get generative AI (such as GPT or Bard) to answer in the way that I wanted.	69.8	70.1	73.3	79.9
I could evaluate the quality of generative AI outputs.	72.0	75.2	76.1	82.6
I could tell the difference between a prompt that was likely to be effective from a prompt that was likely to be ineffective.	66.7	69.8	71.2	81.0
Total Self-Efficacy Mean (standard deviation)	69.5 (17.4)	71.7 (18.0)	73.5 (17.0)	81.2 (11.6)

Table 1. *Students self-reported self-efficacy on various elements of the assignment, reflecting on skills before ("Retro") and after ("Current") the first and last iteration of the generative AI assignment.*

Performance Assignments were graded on three dimensions: Quality of Initial Argument, Quality of Prompt Engineering, and Quality of Response to AI Feedback, each on a 3-point scale. Paired-sample t-tests comparing each student's scores on Assignment 1 (pre-test) with that same student's score on Assignment 5 (post-test) were used to determine statistical significance. Students showed significant improvement on all three dimensions (see Fig. 1): Initial Argument ($\bar{x}_{assignment1} = 2.58$, $\bar{x}_{assignment5} = 2.90$, $t(109) = 6.22$, $p < .001$), Prompt Engineering ($\bar{x}_{assignment1} = 2.34$, $\bar{x}_{assignment5} = 2.89$, $t(109) = 8.80$, $p < .001$), and Response to Feedback ($\bar{x}_{assignment1} = 2.41$, $\bar{x}_{assignment5} = 2.88$, $t(109) = 5.44$, $p < .001$). Student improvement was equivalent to approximately a full standard deviation on each of the dimensions. Students were allowed to skip one assignment, which meant that not all students completed both assignments. This could introduce bias into the analysis if, for example, stronger students completed Assignment 5 than Assignment 1. However, the opposite trend was true – the strongest students in the class had scored so highly on initial assignments that they were in a position to drop the last assignment. Moreover, the analysis was robust to inclusion of all students in the sample, although such an analysis violates the independence assumptions of the statistical tests and should be interpreted accordingly.

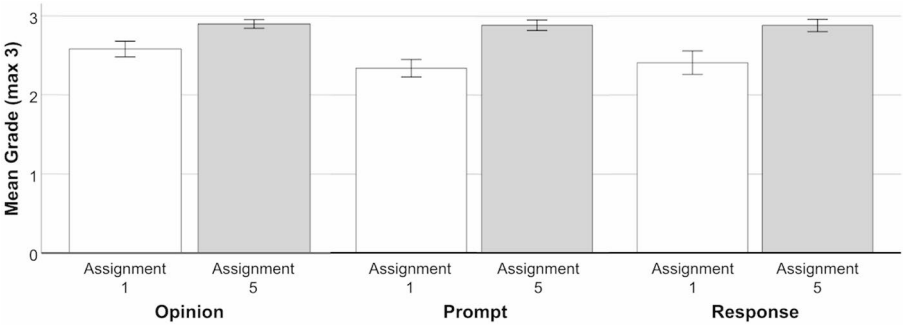


Fig. 1 Performance on the first and last iteration of the assignment

Figure 1. *"Opinion" is the grade students received for their initial argumentative response. "Prompt" is the grade they received for their prompt engineering. "Response" is the grade earned for how well they responded to the critiques generated by the generative AI. Error bars are 95% confidence intervals; in all cases, differences between Assignment 1 and Assignment 5 are significant at the p <.001 level. This figure was created in SPSS and Microsoft Powerpoint.*

Table 2 Examples of Reflection, Rebuttal, and Acceptance

Engagement Type	Quote from Student Essay
Reflection	"I think what GPT did in its counterargument is also effective. It points out the 'core definition' of intelligence to illustrate its main components, exemplifying some of the intelligence capacities, which provides a more straightforward and correlated response to the prompt of "What is Intelligence?"
Reflection	"I think GPT might have misunderstood my response in its third and fourth points of critique…"
Reflection	"ChatGPT's counter arguments provide valuable insights into the need for balanced perspectives…."
Rebuttal	"I don't necessarily think that GPT's example of a music prodigy that does a lackluster job at traditional IQ tests undermines the value of General G. I still hold the belief that said music prodigy would likely demonstrate a higher level of intelligence in other areas that perhaps the IQ tests given to him do not capture, for example, motor skills, spatial awareness, etc."
Rebuttal	"One of ChatGPT's overarching points is that … potential inquiries into intelligence should be handled with care - regulatory frameworks in study design and ethical reporting in the media. However, this point does not fully take into account the forces at play behind the media and general public."
Acceptance	"ChatGPT states 'research in neuroscience has identified structural and functional differences in the brains of individuals, which may predispose them to excel in specific domains.' I agree with this statement because there are differences in people's cognitive abilities that are predetermined by the brain and cannot be improved no matter how much time and effort an individual puts in.
Acceptance	"I concede that I might've been a bit ambitious to say that intelligence was something unchangeable. I wasn't considering the idea of meta-learning, learning to learn, properly."

Engagement with LLM Feedback To assess the extent to which students' responses engaged with the LLM feedback, we recruited two undergraduate research assistants (blind to experimental hypotheses) to code each student's response on three dimensions. First, the coders assessed whether the student **reflected** on the LLM's response by "evaluating the quality of the output, comparing the LLM's output to their own thoughts, or explicitly discussing the LLM's output." On average, the coders found evidence of reflection in 92.7% of responses. Second, the coders assessed whether the student **rebutted** any claims made by the LLM by "arguing against any of the feedback" or "explaining why the LLM's response may not be valid/appropriate." On average, the coders found evidence of rebuttal in 87.8% of responses. Finally, the coders assessed whether each student **accepted** feedback from the LLM by "acknowledging the accuracy or value of the output," or "incorporating it into their own thinking." On average, the coders found evidence of acceptance in 93.6% of responses. Inter-rater reliability was high for each of the three codes (ks = 0.81 −0.89). Examples of reflection, rebuttal, and acceptance are provided in Table 2. These results suggest that the vast majority of students were engaging with the LLM feedback on a deep level and making active decisions about what feedback to accept and what feedback to rebut.

Qualitative Self-Report Feedback In the surveys, students were asked several open-ended questions about their experiences with the generative AI intervention. The first two questions specifically asked students to comment on strengths and weaknesses

of the intervention, respectively, while the third asked students to comment more generally on their experiences. We asked ChatGPT4 to run a sentiment analysis on the collected responses (prompts in the Supplementary Materials). The literature suggests that LLMs can conduct sentiment analyses that are accurate and effective across many domains (Upadhye, 2024), including in the domain of student feedback (Shaikh et al., 2023). Unsurprisingly, in the present data, the sentiment of the responses to the question about strengths of the intervention were positive, and the sentiment of the responses to the question about weaknesses were negative. Encouragingly, for the general feedback, 92.9% of the responses were coded by GPT4 as positively valenced, 7.1% were coded as neutrally valenced, and 0% were coded as negatively valenced. These findings align with and serve to reinforce the results of the affective response data reported above.

As a highly exploratory analysis, we also asked ChatGPT4 to identify themes in the student responses. The literature on how well LLMs can identify themes in qualitative text is contentious (Dai et al., 2023; Zhang et al., 2024), and as such the results should be viewed as preliminary and exploratory and will not be discussed in detail here. However, the complete classification scheme and the relative distribution of responses within that scheme can be found in the Supplementary Materials.

Discussion

Artificial intelligence systems have been used for years to promote content mastery in educational settings. However, the recent development of generative Large Language Models has changed the landscape of how AI can be used in the classroom. LLMs are distinct from older AI models in that they are conversational agents that mimic human language (Radivojevic et al., 2024; Sun & Wang, 2025) and social norms (Horiguchi et al., 2024). As such, interacting with LLMs feels more like having a chat with another human, rather than being taught content by a tutor or teacher. The conversational nature of LLMs presents a unique opportunity for students to use LLMs not only as sources of knowledge, but rather as collaborators in their learning process. Students can use LLMs as listeners to whom they can elaborate their current knowledge, sources of information that can fill gaps in their knowledge, and argumentative (in the good sense) partners who can challenge them to think more deeply about their beliefs (Andrews & Rapp, 2015; Chinn & Clark, 2013; D'Mello et al., 2014; Nussbaum, 2008; Van Blankenstein et al., 2013; Webb, 1991). In other words, LLMs have the potential to act as collaborative learning partners with students.

This collaborative capacity is potentially valuable because the educational literature has consistently highlighted collaborative learning - at least between humans - as a powerful pedagogical tool for facilitating educational outcomes (Kirschner et al., 2011; Terenzini et al., 2001; Van Boxtel et al., 2000). If collaborating with LLMs could produce similar educational benefits as collaborating with humans - without the logistical and interpersonal challenges of organizing teams of students to work together on projects (Lee et al., 2015; North et al., 2000) - then it may be of great value to educators to implement interventions that encourage students to engage in collab-

orative learning with LLMs. For that to happen, however, students must be willing to engage with LLMs in the manner that allows collaborative learning to be effective. However, to date there has been little research on how students interact with LLMs in scaffolded collaborative learning contexts. In this exploratory study, we investigated students' affective and metacognitive responses to a scaffolded LLM assignment, and examined the extent and nature of student engagement with the LLMs.

Students were tasked with using generative AI to improve their argumentative writing skills - such as making claims, providing grounds and warrants, offering qualifiers, and responding to rebuttals (e.g., Nussbaum, 2008; Sampson & Clark, 2009; Toulmin, 2003) - by having LLMs review and critique their argumentative essays. This intervention was designed to emulate typical classroom activities in which students review and critique each others' writing before the assignment is due to the instructor. Across all measures, we found consistently promising results: Over the course of the semester, students became more confident in their abilities, and that confidence was justified by actual improvements in performance. Students became more effective at using LLMs (prompt engineering), more skilled at responding effectively to LLM feedback, and perhaps most importantly, better able to develop arguments even prior to interacting with LLMs (i.e., during their initial draft). While it is difficult to draw causal inferences about the reasons for these improvements, we were able to explore the extent to which the behaviors that previous literature have identified as central for effective collaborative learning were observed in this assignment.

Reflecting back on the nascent literature on LLMs as collaborative learning partners, there were three key concerns about whether LLMs could effectively function in this role. The first concern is whether students would engage with the task and content (which would support learning), or if they would simply offload the assignment to the LLM (which would undermine learning; cf. Hamilton & Benjamin, 2019; Hargis, 2019). We found that over 90% of the assignments showed evidence of active reflection, meaning that students were using AI in a way that pedagogy experts would endorse.

A second possible issue with using LLMs as collaborative learning partners is the question of whether LLMs would be as effective as peers in fostering the disagreement and resolution which is so central to effective collaborative learning (Chin & Osborne, 2010; Gadgil & Nokes-Malach, 2012; Kim et al., 2024; Nussbaum, 2008). Disagreement and debate could fail if students are unable to maintain the necessary skepticism to be willing to refute the LLM (e.g., if the LLM were viewed as an expert), or if the LLM is unwilling to contradict or criticize the student. In the present study, both of these concerns were unfounded: we found that almost 90% of assignments included the student rebutting at least one argument from the LLM, and over 90% of the assignments included the student updating their own response because of criticism they had received from the LLM. Similar to Essel et al. (2024) and Mahapatra (2024), students acted as critical users of the LLMs, capable of highlighting both their strengths and their limitations, suggesting that students were actively engaging with the feedback that they received rather than passively adopting suggestions. Theory suggests that collaborative learning interventions using LLMs would work best if students approach LLMs not as experts or reference tools, but rather as sound-

ing boards for discussion and deliberation, and that is exactly how students interacted with AI in this scaffolded assignment.

The final concern about LLMs as collaborative learning partners is that students may simply dislike working with LLMs and refuse to engage with the assignment, or otherwise find the assignment frustrating or demotivating. We found no evidence for that possibility. Students found the assignments to be challenging, engaging, useful, and enjoyable, and that was true even through the fifth and final iteration of the assignment. Regardless of whether subjective experience with the assignment was measured using open-ended responses or self-reported affective and metacognitive evaluations (measured via Likert scales), students reliably found the assignments to be a positive experience.

In other words, students interacted with the LLMs much like they do with human peers during collaborative learning exercises, and we observed similar learning outcomes as one would expect based on the peer-to-peer collaborative learning literature. These results suggest that students were able to effectively engage in collaborative learning with LLMs.

Limitations

The consistency of these results is highly encouraging, particularly given that the intervention was implemented with real students in an ecologically-valid classroom setting. Not only did the students show actual performance improvements over the course of the semester, but they also had positive attitudes towards the intervention - a critical factor in generating student buy-in. It is also noteworthy that this intervention was implemented in a large class of more than 150 students, supporting the scalability of the intervention with minimal cost. However, the ecological validity of this study came with limitations in internal validity.

First and foremost, there was no control condition: all students in the class completed the assignments in collaboration with LLMs. This was due to logistical, pedagogical, and ethical constraints on the nature of the sample – students within a single class and graded on a single curve could not receive differential treatment in terms of training and assignments. While we observed substantial growth over the semester, there is no guarantee that such growth would not have occurred even in the absence of the intervention. Perhaps other elements of the course, such as the lectures or class discussions or practice with generative AI outside of the class were (partially) responsible for the observed improvements in argumentative writing, prompt engineering, and self-efficacy. Future studies should use randomized controlled designs either in lab settings or across different sections of a course to rule out this challenge. As it stands, the results are suggestive, but one cannot draw strong causal inferences about how much of the observed learning was due specifically to the intervention.

Secondly, there were differences between the first and fifth assignment that could have led to different performance metrics unrelated to actual learning. For example, the first assignment was graded early in the semester and the fifth was graded at the end of the semester. While the teaching assistant who graded the assignments was not aware that this study was occurring, and thus would not have been subject to study-related demand characteristics, he could have been more lenient as the term

progressed due to fatigue or other factors. Similarly, the first assignment could theoretically have been easier than the last assignment, although the assignment prompts were written before the teaching team implemented this specific intervention and were intended to be of similar levels of difficulty. Ideally, future tests of this kind of intervention should counterbalance the order of the prompts, perhaps across sections of a course.

Finally, there are questions about the extent to which these results would generalize to other populations, other disciplines, and other learning goals. This study was run at a highly selective R1 university which is particularly known for its students' work ethic and tech savviness. Other populations may react differently to such a generative AI-focused assignment or may not have the requisite skill set with technology as a foundation to build more advanced skills. It is also possible that the specific content of this class (a psychology-based general education course that fulfills students' "disciplinary perspectives in social sciences" requirement) was particularly amenable to this form of intervention. Whether courses that focus on knowledge retention, programming or computational skills, or which do not involve argumentation-based skills would show the same degree of benefit from this intervention is an open empirical question. Although this intervention might be modified to fit other classroom types or disciplines, it is unclear if such modifications would still yield similar results. Given the great potential for LLMs to inform and improve classroom practices, exploring the effectiveness of LLMs for supporting other learning goals should be a high priority for the learning sciences.

All of these caveats suggest that scholars should attempt to replicate, extend, and generalize these findings before widespread adoption of this sort of intervention is recommended. Nonetheless, the results from this study are robust, remarkably consistent, and quite auspicious in suggesting that when appropriately structured and with scaffolds in place, LLMs can be effectively integrated into classroom assignments as collaborative learning partners At the very least, students find them useful, enjoyable, confidence boosting, and engage with the LLMs in ways that the extant literature suggests would be helpful - all findings that hold regardless of the aforementioned weaknesses in the study design.

Challenges of Using LLMs in the Classroom

Though the results presented here suggest that students can use LLMs to engage in collaborative learning in ways that may augment performance and promote learning, the implementation of generative AI in the classroom is not without risks or challenges. For example, there is a risk that students may outsource formative assignments to LLMs, rather than using the LLMs as tools to facilitate learning, which can undermine learning goals (Hamilton & Benjamin, 2019; Hargis, 2019). While that did not appear to be a problem in this study, it is a concern about which instructors must be constantly aware.

Similarly, supposed learning gains observed while students still have access to generative AI can obscure dependence on the tool (Fan et al., 2024). In the present study, students ostensibly did not use ChatGPT in the development of their initial argument (although compliance to this instructional constraint was not measured),

and yet we saw significant improvement in their ability to articulate their position from the first to the fifth assignment. Although we should be careful not to over-interpret this in light of the fact that students may have used an LLM when they were not supposed to, given that at no point in the class was any academic integrity violation ever detected, the pattern of results suggests that students did not develop an overdependence on the LLMs. Instead, students appeared to develop their own skills to generate effective arguments, even when crafting those arguments without the help of generative AI. Future research should more carefully monitor students to ensure that cheating is not occurring, to strengthen our confidence that students are not developing overdependence.

From a pedagogical perspective, an additional challenge of implementing LLM-supported activities is that students may not begin the intervention with the requisite skills to effectively navigate collaboration with LLMs. As such, instructors must be willing to teach students *how* to collaborate with LLMs and support the intervention appropriately - much like they would support any other activity or intervention, including collaborations with human peers (Salmons, 2023). Critically, the results of this study demonstrated that students report greater self-efficacy in their ability to work with LLMs at the end of the semester than at the beginning of the semester. Although we cannot make causal claims given the lack of a control condition, this result is nonetheless encouraging in that it suggests that, with practice, students can become more comfortable collaborating with LLMs.

Conclusion

Generative AI – including Large Language Model (LLM) chatbots, like ChatGPT – has the potential to provide individualized feedback, instruction, and tutoring to students of all ages. However, it also comes with risks of turning into a crutch and undermining, rather than supporting, student learning. Given the wide range of possible benefits and risks of using generative AI in the classroom, it is critical that we understand strategies that instructors can implement to scaffold students' use of AI to maximize benefits and minimize risks. The present study developed and explored one such strategy. Students independently used LLMs to critique their argumentative writing, but specific instructions were provided to guide students towards engaging with the generative AI in ways that aligned with best practices for collaborative learning (Chin & Osborne, 2010; Hmelo-Silver & Barrows, 2008; Kim et al., 2024; Nussbaum, 2008; Van Boxtel et al., 2000) while also achieving course learning goals. The intervention was also designed to train students on how to use LLMs more effectively so that they could coerce the LLMs to provide more helpful, critical feedback that would improve students' subsequent performance. Indeed, we found that students did use the LLMs in ways that align with best practices for collaborative learning. We hope the present manuscript serves as an early building block towards effectively and equitably leveraging generative AI in the classroom.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40593-025-00521-6.

Acknowledgements The authors would like to thank Yulin Song, Chad "Donut King" Hershock, and the OppLab for assistance with the paper. Special thanks to our coders, Piper Ligotti and Roslyn Randall.

Author Contributions D.M.O. and A.E.C.P. designed the intervention and assessments. D.M.O. ran the study and made bad puns. A.E.C.P. conducted the analysis and generated the tables and figures. T.N.C. conducted the literature review. All authors contributed to the writing and editing of the manuscript.

Funding Open Access funding provided by Carnegie Mellon University. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability The data is available on the Open Science Framework: https://osf.io/74xjb/?view_only=58c61366594d4c96a48854f9ea25b6e8.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alam, A. (2022). Employing adaptive learning and intelligent tutoring robots for virtual classrooms and smart campuses: Reforming education in the age of artificial intelligence. *Advanced Computing and Intelligent Technologies*, *914*, 395–406. https://doi.org/10.1007/978-981-19-2980-9_32

Andrews, J. J., & Rapp, D. N. (2015). Benefits, costs, and challenges of collaboration for learning and memory. *Translational Issues in Psychological Science*, *1*(2), 182–191. https://doi.org/10.1037/tps0000025

Ashktorab, Z., Liao, Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2020). Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW2), 1–20. https://doi.org/10.1145/3415167

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W.H. Freeman and Company.

Baroroh, D. K., Chu, C. H., & Wang, L. (2021). Systematic literature review on augmented reality in smart manufacturing: Collaboration between human and computational intelligence. *Journal Of Manufacturing Systems, 61*, 696–711. https://doi.org/10.1016/j.jmsy.2020.10.017

Carnegie Mellon University (2024). *Diversity, Equity, Inclusion, and Belonging*. https://www.cmu.edu/diversity/tartans/index.html

Cash, T. N., & Oppenheimer, D. M. (2024). Generative chatbots ain't experts: Exploring cognitive and metacognitive limitations that hinder expertise in generative chatbots. *Journal of Applied Research in Memory and Cognition*, *13*(4), 490–494. https://doi.org/10.1037/mac0000202

Cash, T. N., Oppenheimer, D. M., Christie, S., & Devgan, M. (2025). Quantifying uncert-AI-nty: Testing the accuracy of llms' confidence judgments. *Memory & Cognition*. Advance online publication. https://doi.org/10.3758/s13421-025-01755-4

Chen, Y., Fu, Q., Yuan, Y., Wen, Z., Fan, G., Liu, D., Zhang, D., Li, Z., & Xiao, Y. (2023). Hallucination detection: Robustly discerning reliable answers in large language models. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 245*, 255. https://doi.org/10.1145/3583780.3614905

Chin, C., & Osborne, J. (2010). Students' questions and discursive interaction: Their impact on argumentation during collaborative group discussions in science. *Journal of Research in Science Teaching*, *47*(7), 883–908. https://doi.org/10.1002/tea.20385

Chinn, C. A., & Clark, D. B. (2013). Learning through collaborative argumentation. In C. E. Hmelo-Silver, C. A. Chinn, C. Chan, & A. O'Donnell (Eds.), *The international handbook of collaborative learning* (1st ed., pp. 314–332). Routledge/Taylor & Francis Group.

Chinn, C. A., O'donnell, A. M., & Jinks, T. S. (2000). The structure of discourse in collaborative learning. *The Journal of Experimental Education*, *69*(1), 77–97. https://doi.org/10.1080/00220970009600650

Chowdhary, K. R. (2020). Natural Language Processing. In *Fundamentals of Artificial Intelligence* (pp. 603–649). Springer. https://doi.org/10.1007/978-81-322-3972-7_19

Cohen, E. G., Lotan, R. A., Scarloss, B. A., & Arellano, A. R. (1999). Complex instruction: Equity in cooperative learning classrooms. *Theory Into Practice, 38*(2), 80–86. https://doi.org/10.1080/00405849909543836

Congleton, A. R., & Rajaram, S. (2011). The influence of learning methods on collaboration: Prior repeated retrieval enhances retrieval organization, abolishes collaborative inhibition, and promotes post-collaborative memory. *Journal of Experimental Psychology: General*, *140*(4), 535–551. https://doi.org/10.1037/a0024308

Connick, S., & Innes, J. E. (2003). Outcomes of collaborative water policy making: Applying complexity thinking to evaluation. *Journal of Environmental Planning and Management, 46*(2), 177–197. https://doi.org/10.1080/0964056032000070987

Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, *385*(6714), eadq1814. https://doi.org/10.1126/science.adq1814

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29*, 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003

Dai, S. C., Xiong, A., & Ku, L. W. (2023). *LLM-in-the-loop: Leveraging large language model for thematic analysis*. arXiv. https://doi.org/10.48550/arXiv.2310.15100

Desai, S., Dubiel, M., Zargham, N., Mildner, T., & Spillner, L. (2025). Personas evolved: Designing ethical LLM-based conversational agent personalities. *ArXiv*. https://doi.org/10.48550/ArXiv.2502.20513

Erkens, M., & Bodemer, D. (2019). Improving collaborative learning: Guiding knowledge exchange through the provision of information about learning partners and learning contents. *Computers & Education, 128*, 452–472. https://doi.org/10.1016/j.compedu.2018.10.009

Essel, H. B., Vlachopoulos, D., Essuman, A. B., & Amankwa, J. O. (2024). ChatGPT effects on cognitive skills of undergraduate students: Receiving instant responses from AI-based conversational large language models (LLMs). *Computers and Education: Artificial Intelligence, 6*, 100198. https://doi.org/10.1016/j.caeai.2023.100198

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2024). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Felton, M., Garcia-Mila, M., Villarroel, C., & Gilabert, S. (2015). Arguing collaboratively: Argumentative discourse types and their potential for knowledge building. *British Journal of Educational Psychology, 85*(3), 372–386. https://doi.org/10.1111/bjep.12078

Ferretti, R. P., & Graham, S. (2019). Argumentative writing: Theory, assessment, and instruction. *Reading and Writing*, *32*(6), 1345–1357. https://doi.org/10.1007/s11145-019-09950-x

Gabelica, C., De Maeyer, S., & Schippers, M. C. (2022). Taking a free ride: How team learning affects social loafing. *Journal of Educational Psychology, 114*(4), 716–733. https://doi.org/10.1037/edu0000713

Gadgil, S., & Nokes-Malach, T. J. (2012). Overcoming collaborative inhibition through error correction: A classroom experiment. *Applied Cognitive Psychology, 26*(3), 410–420. https://doi.org/10.1002/acp.1843

Google (2024). *Gemini* [Computer software]. https://gemini.google.com/app

Hamilton, K. A., & Benjamin, A. S. (2019). The human-machine extended organism: New roles and responsibilities of human cognition in a digital ecology. *Journal of Applied Research in Memory and Cognition*, *8*(1), 40–45. https://doi.org/10.1016/j.jarmac.2019.01.001

Hargis, M. B. (2019). *Aging, Motivation, and Memory for Important Information* [Doctoral Dissertation, University of California Los Angeles]. https://escholarship.org/uc/item/076154vm

Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction, 26*(1), 48–94. https://doi.org/10.1080/07370000701798495

Hmelo-Silver, C. E., & Chinn, C. A. (2016). Collaborative learning. In L. Corno, & E. M. Anderman (Eds.), *Handbook of educational psychology* (3rd ed., pp. 349–363). Routledge/Taylor & Francis Group.

Hoegl, M., & Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, *12*(4), 435–449. https://doi.org/10.1287/orsc.12.4.435.10635

Hoffmann, M. H. G. (2015). Changing philosophy through technology: Complexity and computer-supported collaborative argument mapping. *Philosophy & Technology, 28*(2), 167–188. https://doi.org/10.1007/s13347-013-0143-6

Horiguchi, I., Yoshida, T., & Ikegami, T. (2024). Evolution of social norms in LLM agents using natural language. *ArXiv*. https://doi.org/10.1016/j.neucom.2025.130193

Hu, T., & Collier, N. (2024). Quantifying the persona effect in LLM simulations. *ArXiv*. https://doi.org/10.48550/ArXiv.2402.10811

Iordanou, K., Kendeous, P., & Beker, K. (2016). Argumentative reasoning. In J. A. Greene, W. A. Sandoval, & I. Bråten (Eds.), *Handbook of epistemic cognition* (1st ed., pp. 39–52). Routledge.

Jonassen, D. H., & Kim, B. (2010). Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research And Development, 58*(4), 439–457. https://doi.org/10.1007/s11423-009-9143-8

Kim, J., Lee, H., & Cho, Y. H. (2022). Learning design to support student-AI collaboration: Perspectives of leading teachers for AI in education. *Education and Information Technologies*, *27*(5), 6069–6104. https://doi.org/10.1007/s10639-021-10831-6

Kim, S., Lin, T. J., Glassman, M., Ha, S. Y., Wen, Z., Nagpal, M., Cash, T. N., & Kraatz, E. (2024). Linking knowledge justification with peers to the learning of social perspective taking. *Journal of Moral Education*, *53*(2), 321–341. https://doi.org/10.1080/03057240.2023.2185597

Kirschner, F., Paas, F., Kirschner, P. A., & Janssen, J. (2011). Differential effects of problem-solving demands on individual and collaborative learning outcomes. *Learning and Instruction*, *21*(4), 587–599. https://doi.org/10.1016/j.learninstruc.2011.01.001

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30–43.

Kuhn, D. (1991). *The Skills of Argument* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511571350

Kuhn, D., & Crowell, A. (2011). Dialogic argumentation as a vehicle for developing young adolescents' thinking. *Psychological Science*, *22*(4), 545–552. https://doi.org/10.1177/0956797611402512

Laal, M., & Laal, M. (2012). Collaborative learning: What is it? *Procedia - Social And Behavioral Sciences, 31*, 491–495. https://doi.org/10.1016/j.sbspro.2011.12.092

Larrain, A., Singer, V., Strasser, K., Howe, C., López, P., Pinochet, J., Moran, C., Sánchez, Á., Silva, M., & Villavicencio, C. (2021). Argumentation skills mediate the effect of peer argumentation on content knowledge in middle-school students. *Journal Of Educational Psychology, 113*(4), 736–753. https://doi.org/10.1037/edu0000619

Latham, A., Crockett, K., & McLean, D. (2014). An adaptation algorithm for an intelligent natural language tutoring system. *Computers & Education, 71*, 97–110. https://doi.org/10.1016/j.compedu.2013.09.014

Lazakidou, G., & Retalis, S. (2010). Using computer supported collaborative learning strategies for helping students acquire self-regulated problem-solving skills in mathematics. *Computers & Education, 54*(1), 3–13. https://doi.org/10.1016/j.compedu.2009.02.020

Lee, D., Huh, Y., & Reigeluth, C. M. (2015). Collaboration, intragroup conflict, and social skills in project-based learning. *Instructional Science*, *43*(5), 561–590. https://doi.org/10.1007/s11251-015-9348-7

Lee, M., Gero, K. I., Chung, J. J. Y., Shum, S. B., Raheja, V., Shen, H., Venugopalan, S., Wambsganss, T., Zhou, D., Alghamdi, E. A., August, T., Bhat, A., Choksi, M. Z., Dutta, S., Guo, J. L. C., Hoque, M. N., Kim, Y., Knight, S., Neshaei, S. P., & Siangliulue, P. (2024). A design space for intelligent and interactive writing assistants. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–35. https://doi.org/10.1145/3613904.3642697

Lewis, J. M., Ross, S., & Holden, T. (2012). The how and why of academic collaboration: Disciplinary differences and policy implications. *Higher Education*, *64*(5), 693–708. https://doi.org/10.1007/s10734-012-9521-8

Lin, T. J., Ha, S. Y., Li, W. T., Chiu, Y. J., Hong, Y. R., & Tsai, C. C. (2019). Effects of collaborative small-group discussions on early adolescents' social reasoning. *Reading and Writing*, *32*(9), 2223–2249. https://doi.org/10.1007/s11145-019-09946-7

Lin, T. J., Kraatz, E., Ha, S. Y., Hsieh, M., Glassman, M., Nagpal, M., Sallade, R., & Shin, S. (2022). Shaping classroom social experiences through collaborative small-group discussions. *British Journal of Educational Psychology, 92*(1), 131–154. https://doi.org/10.1111/bjep.12442

Lipponen, L., & Lallimo, J. (2004). Assessing applications for collaboration: From collaboratively usable applications to collaborative technology. *British Journal of Educational Technology*, *35*(4), 433–442. https://doi.org/10.1111/j.0007-1013.2004.00402.x

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments*, *11*(1), 9. https://doi.org/10.1186/s40561-024-00295-9

Maloney, J., & Simon, S. (2006). Mapping children's discussions of evidence in science to assess collaboration and argumentation. *International Journal of Science Education*, *28*(15), 1817–1841. https://doi.org/10.1080/09500690600855419

Melis, A. P. (2013). The evolutionary roots of human collaboration: Coordination and sharing of resources. *Annals of the New York Academy of Sciences, 1299*(1), 68–76. https://doi.org/10.1111/nyas.12263

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700. https://doi.org/10.1016/j.tics.2016.07.001

Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. *Nature Machine Intelligence, 5*(4), 333–334. https://doi.org/10.1038/s42256-023-00644-2

Mirzakhmedova, N., Gohsen, M., Chang, C. H., & Stein, B. (2024). *Are Large Language Models Reliable Argument Quality Annotators?* arXiv. https://doi.org/10.48550/ARXIV.2404.09696

Münchow, H., Richter, T., von der Mühlen, S., & Schmid, S. (2019). The ability to evaluate arguments in scientific texts: Measurement, cognitive processes, Nomological network, and relevance for academic success at university. *British Journal of Educational Psychology*, *89*(3), 501–523. https://doi.org/10.1111/bjep.12298

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association, 18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Näykki, P., Järvelä, S., Kirschner, P. A., & Järvenoja, H. (2014). Socio-emotional conflict in collaborative learning—A process-oriented case study in a higher education context. *International Journal of Educational Research*, *68*, 1–14. https://doi.org/10.1016/j.ijer.2014.07.001

Nokes-Malach, T. J., Richey, J. E., & Gadgil, S. (2015). When is it better to learn together? Insights from research on collaborative learning. *Educational Psychology Review*, *27*(4), 645–656. https://doi.org/10.1007/s10648-015-9312-8

North, A. C., Linley, P. A., & Hargreaves, D. J. (2000). Social loafing in a co-operative classroom task. *Educational Psychology*, *20*(4), 389–392. https://doi.org/10.1080/01443410020016635

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Nussbaum, M. E. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology*, *33*(3), 345–359. https://doi.org/10.1016/j.cedpsych.2008.06.001

OpenAI (2024). *ChatGPT* [Computer software]. https://chat.openai.com/

Osborne, J. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, *328*(5977), 463–466. https://doi.org/10.1126/science.1183944

Pajares, F., Hartley, J., & Valiante, G. (2001). Response format in writing self-efficacy assessment: Greater discrimination increases prediction. *Measurement and Evaluation in Counseling and Development*, *33*(4), 214–221. https://doi.org/10.1080/07481756.2001.12069012

Preiss, D. D., Castillo, J. C., Grigorenko, E. L., & Manzi, J. (2013). Argumentative writing and academic achievement: A longitudinal study. *Learning and Individual Differences*, *28*, 204–211. https://doi.org/10.1016/j.lindif.2012.12.013

Radivojevic, K., Chou, M., Badillo-Urquiola, K., & Brenner, P. (2024). Human perception of LLM-generated text content in social media environments. *ArXiv*. https://doi.org/10.48550/ArXiv.2409.06653

Rajaram, S., & Pereira-Pasarin, L. P. (2010). Collaborative memory: Cognitive research and theory. *Perspectives on Psychological Science*, *5*(6), 649–663. https://doi.org/10.1177/1745691610388763

Retnowati, E., Ayres, P., & Sweller, J. (2018). Collaborative learning effects when students have complete or incomplete knowledge. *Applied Cognitive Psychology*, *32*(6), 681–692. https://doi.org/10.1002/acp.3444

Roose, K. (2023). *March 8). How does ChatGPT really*. work? The New York Times.

Salmons, J. (2023). *Learning to collaborate, collaborating to learn: Engaging students in the classroom and online* (1st ed.). Routledge. https://doi.org/10.4324/9781003445708

Sampson, V., & Clark, D. (2009). The impact of collaboration on the outcomes of scientific argumentation. *Science Education*, *93*(3), 448–484. https://doi.org/10.1002/sce.20306

Sawyer, R. K., & DeZutter, S. (2009). Distributed creativity: How collective creations emerge from collaboration. *Psychology of Aesthetics, Creativity, and the Arts, 3*(2), 81–92. https://doi.org/10.1037/a0013282

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning, 5*(1), 43–102. https://doi.org/10.1007/s11412-009-9080-x

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570–582. https://doi.org/10.1086/269282

Schwarz, N., Grayson, C. E., & Knäuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10*(2), 177–183. https://doi.org/10.1093/ijpor/10.2.177

Shaikh, S., Daudpota, S. M., Yayilgan, S. Y., & Sindhu, S. (2023, December). Exploring the potential of large-language models (LLMs) for student feedback sentiment analysis. *International Conference on Frontiers of Information Technology*, 214–219. https://doi.org/10.1109/FIT60620.2023.00047

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: Systematic review. *Journal of Medical Internet Research, 26*, e52597. https://doi.org/10.2196/52597

Sowa, K., Przegalinska, A., & Ciechanowski, L. (2021). Cobots in knowledge work. *Journal of Business Research*, *125*, 135–142. https://doi.org/10.1016/j.jbusres.2020.11.038

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, *57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Sun, Y., & Wang, T. (2025). Be friendly, not friends: How LLM sycophancy shapes user trust. *ArXiv* https://doi.org/10.48550/ArXiv.2502.10844

Sun, J., Anderson, R. C., Lin, T. J., Morris, J. A., Miller, B. W., Ma, S., Nguyen-Jahiel, T., K., & Scott, T. (2022). Children's engagement during collaborative learning and direct instruction through the lens of participant structure. *Contemporary Educational Psychology*, *69*, 102061. https://doi.org/10.1016/j.cedpsych.2022.102061

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning And Individual Differences, 61*, 136–150. https://doi.org/10.1016/j.lindif.2017.11.015

Terenzini, P. T., Cabrera, A. F., Colbeck, C. L., Parente, J. M., & Bjorklund, S. A. (2001). Collaborative learning vs. lecture/discussion: Students' reported learning gains. *Journal Of Engineering Education, 90*(1), 123–130. https://doi.org/10.1002/j.2168-9830.2001.tb00579.x

Tompson, G. H., & Dass, P. (2000). Improving students' self-efficacy in strategic management: The relative impact of cases and simulations. Simulation & Gaming, 31(1), 22–41.

Toulmin, S. E. (2003). *The Uses of Argument* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Upadhye, A. (2024). Sentiment analysis using large language models: Methodologies, applications, and challenges. *International Journal of Computer Applications, 186*, 30–34. https://doi.org/10.5120/ijca2024923625

Usdan, J., Connell Pensky, A. E., & Chang, H. (2024). Generative ai's impact on graduate student writing productivity and quality. *SSRN*. https://doi.org/10.2139/ssrn.4941022

Van Blankenstein, F. M., Dolmans, D. H. J. M., Van Der Vleuten, C. P. M., & Schmidt, H. G. (2013). Relevant prior knowledge moderates the effect of elaboration during small group discussion on academic achievement. *Instructional Science, 41*(4), 729–744. https://doi.org/10.1007/s11251-012-9252-3

Van Boxtel, C., Van Der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction, 10*(4), 311–330. https://doi.org/10.1016/S0959-4752(00)00002-5

Vidergor, H. E., & Ben-Amram, P. (2020). Khan academy effectiveness: The case of math secondary students' perceptions. *Computers & Education*, *157*, 103985. https://doi.org/10.1016/j.compedu.2020.103985

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., & Leimeister, J. M. (2020). AL: An adaptive learning support system for argumentation skills. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376732

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities And Social Sciences Communications, 12*, 621. https://doi.org/10.1057/s41599-025-04787-y

Wang, L., Chen, X., Wang, C., Xu, L., Shadiev, R., & Li, Y. (2024). ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study. *Thinking Skills and Creativity, 51*, Article 101440. https://doi.org/10.1016/j.tsc.2023.101440

Wang, Q., Fu, Y., Cao, Y., Wang, S., Tian, Z., & Ding, L. (2025a). Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing, 639*, 130193. https://doi.org/10.1016/j.neucom.2025.130193

Wang, Y., He, B., Chen, X., & Sun, L. (2025b). Can LLMs Clarify? Investigation and Enhancement of Large Language Models on Argument Claim Optimization. *Proceedings of the 31st International Conference on Computational Linguistics*. https://aclanthology.org/2025.coling-main.273/

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education*, *22*(5), 366. https://doi.org/10.2307/749186

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education, 46*(1), 71–95. https://doi.org/10.1016/j.compedu.2005.04.003

Weise, K., & Metz, C. (2023, May 9). When A.I. chatbots hallucinate. *The New York Times*. https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html

Wen, Z., Lin, T.-J., Glassman, M., Kim, S., Tilak, S., Nagpal, M., & Ha, S. Y. (2023). The development of early adolescents' social perspective taking through small-group discussion. *The Journal of Early Adolescence, 43*(9), 1129–1163. https://doi.org/10.1177/02724316221149630

Wolfram, S. (2023). *What is ChatGPT doing: … and what does it work?* Wolfram Media, Inc.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zhang, M. J. Q., & Choi, E. (2023). *Clarify when necessary: Resolving ambiguity through interaction with LMs*. arXiv. https://doi.org/10.48550/arXiv.2311.09469

Zhang, P., & Tur, G. (2024). A systematic review of ChatGPT use in K-12 education. *European Journal of Education, 59*(2), Article e12599. https://doi.org/10.1111/ejed.12599

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2024). Redefining qualitative analysis in the AI era: Using ChatGPT for efficient thematic analysis. *ArXiv*. https://doi.org/10.48550/ArXiv.2309.10771

Zhao, H., & Hawkins, R. D. (2025). *Comparing human and LLM politeness strategies in free production*. arXiv. https://doi.org/10.48550/arXiv.2506.09391

Zhao, Y., Zhang, R., Li, W., & Ling, L. (2025). Assessing and understanding creativity in large language models. *Machine Intelligence Research, 22*, 417–436. https://doi.org/10.1007/s11633-025-1546-4

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.