The use of large language models as scaffolds for proleptic reasoning

Olya Kudina, Delft University of Technology
Brian Ballsun-Stanton, Macquarie University
Mark Alfano, Macquarie University

Abstract

This paper examines the potential educational uses of chat based Large Language Models (LLMs), moving past initial hype and skepticism. Although LLM outputs often evoke fascination and resemble human writing, they are unpredictable and must be used with discernment. Several metaphors—like calculators, cars, and drunk tutors—highlight distinct models for student interactions with LLMs, which we explore in the paper. We suggest that LLMs hold a potential in students' learning by fostering proleptic reasoning through scaffolding, i.e., presenting a technological accompaniment in anticipating and responding to potential objections to arguments. Here, the technical limitations of LLMs can be reframed as beneficial when fostering anticipatory reasoning. Whether their outputs are accurate or not, evaluating them stimulates learning. LLMs require students to critically engage, emphasizing analytical thinking over mere memorization. This interaction helps solidify knowledge. Additionally, we explore how engaging with LLMs can prepare students for constructive collective discussions and provide first steps in addressing epistemic injustices by highlighting potential research blind spots. Thus, while acknowledging the sociopolitical and ethical complexities of using LLMs in education, we suggest that when used in an informed way, they can promote critical thinking through anticipatory reasoning.

Keywords: generative pre-trained transformer (GPT), large language model (LLM), proleptic reasoning, pedagogy

Acknowledgments
Olya Kudina wants to acknowledge the support of TU Delft AI Lab's program, particularly the AI DeMoS Lab, in working on this paper.

Introduction

The profusion of Large Language Models (LLMs), such as OpenAI's GPT-4 and its dialogue-based interface ChatGPT, has caused much speculation about their potential to revolutionize education and the nature of learning (Fuchs 2023; Firat 2023; Yu 2023). These language models can generate extensive outputs that resemble human-written text in a matter of seconds, answering questions, translating between languages, and even providing explanations of complex topics, some sound, others unsound. Their output seems credible to many users, resembles human-generated prose, and at first glance requires little effort from a user to generate. It is perhaps unsurprising that soon after the introduction of ChatGPT into society, students and academics around the globe started to experiment with this interactive technology (Ceres, 2023; Howell, 2023; Mollick, 2022), which is always available to anyone with an internet connection and can provide individualized outputs. In this paper, we attend to a more effective technique for the use of LLMs by students, though we of course recognize that they could also potentially be used by teachers and graders.

Parallel to students' exploration of ChatGPT and initial enthusiasm about it, educators have struggled to come up with an appropriate response to the proliferation of LLMs (Ceres and Hoover, 2023). Some worry that they facilitate plagiarism and other academic misconduct. Others are concerned about LLMs' potential to disrupt the conventional learning processes and assessment mechanisms. Still others fear that students will come to think of it as a sort of oracle, reducing their inclination and ability to engage in critical thinking. However, like all classroom disrupting technologies – calculators, PowerPoint, and word processors – effective use of these tools requires updating our techniques and assessments to achieve our pedagogical goals. Setting aside the philosophy of technology, these LLMs are deceptively powerful and are offered in a "chatbot" interface without effective error messages or a strong relationship to facts. The effective use of these tools is a skill like any other, once we see past the perception of having a "conversation." Without discounting the concerns about the appropriate use of technology, in this paper we will explore some positive uses of ChatGPT and other LLMs as scaffolds for proleptic reasoning: the skill of anticipating reactions to arguments (either one's own or another's) and thereby reasoning more thoroughly about a question or problem (Clauss, 2007). Additionally, even though LLMs may be usefully employed by many actors in the education setting, e.g., researchers and teachers, in this paper, we will study the impact of LLMs on the learning process and as such, will focus on the students.

With the right series of "prompts" – conversations more akin to programming or coding – a chat-based LLM can assume the role of a detailed line editor, an assistant smoothing out disfluences of writing English as a second language or helping to overcome the writer's block (Mollick 2023). While the potential to use LLMs to *get started* in the writing process is interesting, we think that there is more value to be wrung from this resource at later stages of the research and writing process, as we explain below. To be clear, poor use of these tools are poor academic practice, demonstrating a failure of judgement and process. Students in Ballsun-Stanton's classes who have used it for higher quality outputs report spending *more*, rather than less time on the assessments. A student Ballsun-Stanton's class on the Techniques of AI in 2024 reflected on the utility, showing how they spent extra time editing:

> Using AI to review assignments against marking criteria has been particularly useful. For instance, when I submitted my research plan along with the marking criteria, Claude provided valuable suggestions for improvement, though I should have asked how I could integrate them into my work. However, this journey has also highlighted the importance of critical thinking when interacting with AI. A recurring theme in our peer discussions was the need to critically evaluate AI outputs.

It is this prompt iteration and evaluation process, on top of the extra editing work which causes effective students to spend more time on high-quality LLM-enhanced outputs. Without this attention to detail, a student's single prompt with unreviewed output will produce what Simon Willison calls "slop" and a sub-par academic output (Willison, 2024).

In articulating our argument, the article proceeds as follows. We first explore various metaphors that have been proposed for understanding LLMs. Next, we explore some promising pedagogical uses of LLMs in the context of scaffolding proleptic reasoning. These uses draw on Clark's (2002) idea that looping our cognition through technological devices and processes is a promising way to improve and refine our cognition. Drawing on philosophical sources from Plato to John Stuart Mill, Harriet Taylor Mill, and Charles Darwin, we contend that LLMs can help students to generate, consider, and respond to potential – both counterarguments to their own conclusions and to those that others might reach. This is an essential part of university education in philosophy, the humanities, and more broadly. It will turn out that the seeming drawbacks of LLMs (e.g., their sometimes unreliability and proneness to engage in "hallucination" or confabulation) are actually benefits in the context of proleptic reasoning and the critical engagement of students on argumentative and tool levels. We also show that LLMs are sometimes useful discovery tools for constructing a more broadly inclusive bibliography. In disciplines such as philosophy where women and non-white contributors are very often ignored or under-cited, LLMs may help students to find sources that they (and, in many cases, their teachers) would otherwise neglect. We do not think that there are only profitable uses of LLMs in education, but we aim to show that there are at least a few positive prospects next to the mounting public concerns.

Metaphors for LLMs

Accompanying this appropriation process is the emergence of several metaphors as a common repertoire upon which one draws when trying to explain either the role, benefits, or shortcomings of ChatGPT and other LLMs in higher education. Some of the ChatGPT metaphors refer to it as a calculator (Bonger et al., 2023)[1], a car with assistive technologies (Morimoto, 2023), or as a drunk tutor.[2] Scrutinizing these metaphors may offer insights into the nature of hopes and worries that the educators have regarding the introduction of LLMs.

---

[1] See also https://simonwillison.net/2023/Apr/2/calculator-for-words/, accessed 28 September 2023.

[2] This metaphor has been proposed by Brian Ballsun-Stanton as part of his Large Language Models workshop. Slides are available in an Open Science Framework repository (url = < https://osf.io/rd24y/, accessed 24 August 2023 >.

Perhaps most frequently, chat-based LLMs are compared to a calculator in an attempt to justify their inevitable societal adoption and to suggest the need to appropriately review learning goals and skillsets in our teaching (Bonger et al., 2023). The reasoning behind this metaphor is delegation of the complex time-consuming tasks to a machine: just as a calculator can quickly solve tedious arithmetic problems that have little intrinsic value, so, one might think, can LLMs alleviate information generation at scale. The assumption here is that LLMs can perform as efficiently and accurately as a calculator. Another important difference is that a calculator will always provide the same output (e.g., '15') to the same input (e.g., '3x5'). Because of the underlying architecture of LLMs, their outputs are only stochastically predictable and may be surprising and even harmful, as in a recent case where an LLM suggested ways to produce chlorine gas when asked for cocktail recipes (McClure, 2023). Another trope inherent to the calculator metaphor is that it is only an addition to human reasoning and can never replace it: just as the basic knowledge of mathematics and of how to operate a calculator are required for using it effectively, so do students need significant knowledge of the subject that they are querying LLMs on and of the workings of LLMs themselves (e.g., their limitations and opportunities, not full technical details) to benefit from their use. Underlying the calculator metaphor is the idea that an LLM is a tool, not a teacher. There is a different problem also addressed using the calculator metaphor. Willison (2023) uses the term "a calculator for words" to address a misconception, saying "One of the most pervasive mistakes I see people using with large language model tools like ChatGPT is trying to use them as a search engine. … [Their appropriate use] is reflected in their name: a "language model" implies that they are tools for working with language. That's what they've been trained to do, and it's language manipulation where they truly excel." Thus, just as a calculator returns 15 from the input "5", "multiply", "3", so too can ChatGPT produce a proleptic response based on one's prompts. Therefore, not only does this tool require careful input, but it also requires us to change our relationship with "truth-producing" tools on the internet.

The use of the car metaphor (e.g. Morimoto, 2023; Shinde, 2023) when related to education positions LLMs as a support technology that helps to achieve a certain end but leaves the responsibility on the student. A car facilitates reaching a certain destination accompanied by features, meant to make driving a more accurate and overall effective process, for instance, with a parking assist technology to help the driver with special maneuvers. LLMs appear on this reading as a facilitator of a learning process, helping to structure and guide it, full of easy-to-access tips, summaries, and being a source of ideas. The car metaphor also underscores the crucial conditions and implications of choosing to use the technology. A driver must first obtain a license and learn to navigate traffic without an assistive technology, and even with the license, experience, and skills, always remains the responsible one, alert to the environment. Similarly, users of LLMs need to first acquire fundamental research, writing, and critical thinking skills before using this technology in their learning practices, and when using it, the students remain responsible for any integration of LLMs' results into their work.

Ballsun-Stanton uses the analogy of "Think of ChatGPT like a drunk tutor: it is always authoritative and usually correct." In his workshop, one of the fundamental themes is that the chatbot's register never reflects its confidence in what it is saying. The drunk tutor metaphor also acknowledges the human knowledge embedded in LLMs while alerting educators and students to the disconnect between patterns of knowledge, facts, and the presentation of knowledge. Because these tools cannot engage with "facts" as a category of knowledge, they are never able to self-

assess the accuracy of their output or represent useful uncertainty. This is the same as a tutor stumbling into class and saying the first thing that comes to mind: they always sound like they know what they're talking about – even if they have no factual basis for their assertions at the time. The metaphor highlights limitations of relying uncritically on the use of ChatGPT. While a drunk tutor or an LLM may provide vast amounts of text, their relevance and credibility must always be questioned. The drunk tutor metaphor also highlights the dangers of anthropomorphizing technologies such as ChatGPT, misconceiving machines that are arguably incapable of induction, discernment, and judgment as knowledgeable and accountable experts producing facts and truths. Just as students would need to question the generous advice of their drunk tutor even if it is sometimes or often brilliant, so do they need to be critical and be able to challenge the output generated by ChatGPT. Having ChatGPT as a helpful hand in the learning process may not be a bad thing so long as its use remains informed and critical.

The use of these metaphors regarding chat-based LLMs reflects the uncertainty as to whether or how they should be formally integrated in the higher education processes. Until now, the official reaction to LLMs in the higher education sector has mirrored a general public perplexity regarding their value, ranging from banning their use completely (e.g. the USA [Johnson, 2023]) to temporarily banning their use until figuring out how exactly it fits in the national laws, e.g. to privacy and data protection (e.g. Italy [McCallum, 2023] to allowing them but coming up with ad hoc committees tasked with developing university-wide or discipline-specific guidelines on their responsible use (e.g. the Netherlands [TU Delft, 2023]), to individual academics experimenting with their use in their classes (e.g., the work of Ethan Mollick on his Blog "One Useful Thing" (2022, 2023)), to silently condoning their use until further notice. In this article, we explore the potential value of LLMs in higher education, especially the humanities, taking into account the perspectives of both students and teachers, and taking philosophy as the primary disciplinary background.

The main idea that we would like to propose is that LLMs, like almost all technologies, are best used by people with a lot of background knowledge, both of their respective domain and of the specific workings of LLMs. Even though LLMs may be useful in exploring new areas and acquiring new skills, we contend that they are most valuable for learners who have already put in a lot of work. Throughout the article, we will explore both the premise and the implications of this claim. In particular, we argue that teaching with and through LLMs is not only possible but even desirable *in the right contexts*. The effective use of LLMs requires from students solid background knowledge and a reflective attitude both to this technology and the output it generates. Just as when talking to a drunk tutor, the students need to be aware that they have to question and validate the output of ChatGPT. Just as it would be pedagogically criticizable to teach students to use calculators without first teaching them mental and pencil-and-paper addition, subtraction, multiplication, and division, so it would be pedagogically criticizable to teach students to use ChatGPT without first teaching them to understand a literature and write critical essays about it. But this does not mean that calculator use should always be forbidden, nor does it mean that LLMs have no place in the crafting of prose. What it does implies, however, is a fundamental reflection on the value and goals of education and the kind of skills we as educators want the students to master.

Scaffolding proleptic reasoning

One skill with which students in philosophy, the humanities, and across the university routinely struggle is proleptic reasoning, by which we mean the anticipation, charitable articulation, and response to potential objections in the form of counterexamples, counterarguments, and so on (Clauss, 2007). Proleptic reasoning is typically social. I offer an argument or make a claim; you challenge it; I respond; perhaps you respond to my response and I respond to your response. Or, you make an argument or claim; I challenge it; you respond; perhaps I respond to your response and you respond to my response. Alternatively, I could consider making an argument or claim and ask myself whether and how you might challenge me, then revise my thinking in advance. You could do the same. Because of its iterative nature, proleptic reasoning makes arguments stronger and more persuasive. It also potentially upgrades true beliefs to knowledge or even understanding. Additionally, it can foster civic engagement. And it can help students to overcome the pervasive problem of confirmation bias.

In this section, we argue that the use of LLMs can support the development of the skill of proleptic reasoning, which can then be translated into social interactions in dialogue, debate, and group inquiry. We do so both by way of theoretical argument and by using a taxonomy of types of proleptic reasoning. As we just pointed out, proleptic reasoning is often social, but it can also be done solo, for instance when brainstorming. Likewise, proleptic reasoning can be diachronic and in-the-moment, as the two initial examples show, but it can also be prospective, as the later examples show. We introduce a third dimension on which proleptic reasoning can vary: whether and how it is technologically scaffolded. We will primarily use examples where it is scaffolded by the use of LLMs; the examples come from a class taught recently by one of the co-authors.

Originating in attempts to understand the acquisition of cognitive abilities in developmental psychology (Wood et al., 1976), the notion of 'scaffolding' has received considerable attention in philosophy of mind and cognition. In contrast to internalist assumptions about the brain-bound realization of mental phenomena (e.g., Fodor, 1980; Adams & Aizawa, 2001), proponents of scaffolding hold that mental processes are often causally influenced by the agent's interaction with environmental resources, e.g., other agents, artefacts, and technological devices (for an overview, see Varga, 2019). Much philosophical research has proceeded by investigating key aspects of cognitive scaffolding (e.g., Clark, 1997; Sterelny, 2010). In what follows, we argue that LLMs can be used as scaffolds for proleptic reasoning.

> In the *Meno* (Cooper, 1997), Plato famously puts these words in the mouth of Socrates: True opinions are a fine thing and do all sorts of good so long as they stay in their place, but they will not stay long. They run away from a man's mind; so they are not worth much until you tether them by working out a reason [....] Once they are tied down, they become knowledge, and are stable. That is why knowledge is something more valuable than right opinion. What distinguishes the one from the other is the tether.

There have been centuries of interpretations of what exactly the character of Socrates means here, and we are not Plato scholars. Therefore, we will just stipulate what we understand by Socrates' reference to the tether in this passage. As we understand it, what distinguishes knowledge (or at least discursive knowledge) from mere true belief and makes such knowledge more valuable than mere true belief is that the person has acquired their true belief through the exercise of epistemic agency or virtue (e.g., Greco 2009). In many cases, especially cases involving linguistic reasoning and exchange, this will supply her with Plato's tether because they

iteratively work out a justification for the true belief that they holds.[3] This in turn means that when they encounter endogenous doubt or are challenged by another epistemic agent, they have an available response that enables them to hold onto their true belief. Without a tether, a reason, they might be hard-pressed to persevere in their true belief.

One good way to acquire a tether is to engage in proleptic reasoning, that is, to engage in the anticipation, charitable articulation, and response to potential objections in the form of counterexamples, counterarguments, and so on. If you've already thought through some of the main reasons that someone might think that you are wrong and have a ready response, you are well-positioned to respond in the case of an actual challenge. This is where the scaffolding of LLMs may come in handy. (Many of us will be familiar with the strategy of including extra slides at the end of a presentation in case a particular objection is voiced, or anticipating the sometimes-uncharitable objections of reviewer 2.)

For example, in the 2024 Techniques of AI unit, one student – looking for arguments around AI's participation in art, queried Claude 3.5 sonnet with the following user prompt (system prompt not detailed here):

> ## User
> Hello, you are going to help me consider a number of perspectives today. All of these perspectives attempt to answer the question
> "is AI art, real art?"
> The perspectives we will be considering are
> 1. Yes it is realy art. It is the intent expressed by typing a prompt or loading the window, is all that is required to make 'art'
> 2. Yes it is art becuase AI is just a tool, in the same way a paintbrush is.
> 3. AI Art might be real art, if the user puts enough time and energy into it
> 4. No AI art is not real art because it requires no skill
> 5. No AI art is not real art because its just not
> Please provide me with arguments for and against all of these perspectives, with sources

The student already had a debate in mind, plus the parameters of their intended position, and their intention with this being the *start* of their conversation is to explore the possible arguments and rebuttals across the range of their already-researched responses to the real art question.

In addition, when a student is considering an argument for a particular conclusion, they could feed their prose into an LLM and ask it to generate a handful of counterarguments. They could then craft responses to these objections. Here, a Macquarie University student was using techniques they learned in Ballsun-Stanton's Techniques of AI unit to improve their assessments for other units. In the middle of their conversation, after they tested many of their arguments out against the LLM interlocutor, they said:

> [H]ere's my revised essay. please mark it harshly against the rubric (do not mark referencing as I have used Zotero for this): ... have I addressed the "political factors" section of the question? PLEASE be honest!!! I am so stressed out because I am tired and this is due.

Here, Claude 3.5 Sonnet then breaks down their revised essay (using the context of their earlier conversation, as well as the essay and rubric) and responds both with a breakdown of strengths and what Claude says: "What could be stronger." Here, the student has explicitly asked for the weaknesses of their argumentation (albeit in an informal tone). They then use Claude's responses to improve their own arguments in an iterative sequence. In this unit, while we did not cover proleptic prompting approaches due to time constraints, the students did learn how to test and

[3] For background and a full reference list see Pritchard et al. (2022).

improve their own writing. With more structure or more time teaching, a more explicit proleptic approach would provide better scaffolding and guidance to students who want to improve their research claims.

For example, as part of our worked example,[4] we opened our conversation with "Hi ChatGPT. Today we will be working on editing a paper that I'm working on. Our goal will be to engage in "proleptic reasoning" and I would like you to take the role of the devil's advocate here. Before we begin though, can you give me a definition of what proleptic reasoning is, and then let's come up with a checklist to apply to the paper. We also need to figure out a way of isolating out our arguments from the text, so that should be part of our checklist.". This prompt establishes a role for ChatGPT to prefer a register and tone from, causes it to define the term, and leads it to create a checklist. By engaging in this "chain of thought" prompting (checklist creation as functional decomposition of the task), we then scaffold the context window. This technique only works, however, if the student needs to engage with, and critically reflect upon the output of the LLM. If the student treats the suggested edits of the LLM as *prima facie* true, there will be confabulations along with a failure to engage with the material. However, by requiring students to annotate their chatlog along with the assignment, we then create multiple opportunities to state tethers and test their utility against an "other's" words.

When prompted, o1's best response was: "3. Address Counterarguments Proleptically ... Here is where you could weave your responses to the novel objections into your essay without rewriting your existing text. You would add clarifications or footnotes, for example, to preempt these criticisms. ... You could specify that "luck" sometimes masks underlying character differences, but a completely untempted individual might be saintly—or might simply lack external pressures." 3.5 Sonnet's best response included: "Vulnerability Analysis Phase ... |W3| There's an unexamined assumption that temptation is necessary for developing executive virtues ... {Counter3} Empirical psychology might suggest that repeatedly resisting temptation depletes willpower, making future resistance harder, not easier." which helps our theoretical student trace their logical premises, vulnerabilities, and assumptions.

The drunk tutor metaphor is instructive here. The responses generated by ChatGPT or Claude might be decisive counterarguments, in which case the student would need to revise their argument. However, even though the arguments sound authoritative and with the register of someone confident in the accuracy of what they are saying, they might also be flawed in various ways, including through hallucination or confabulation. The student would need to exercise their critical reasoning capacities and engage in further original research in order to arrive at a verdict on each of the generated objections. In so doing, they might acquire the tether that Socrates refers to in the *Meno*.

---

[4] Our effective worked example used OpenAI's ChatGPT o1: https://osf.io/gm3uq, a system prompt, https://osf.io/qzxkh, and Claude 3.5 Sonnet's response via Perplexity's interface: https://osf.io/fex3j. This took three attempts, with some prompt editing to make sure that the model followed the checklist we had it establish. This choice of model, and prompt editing is an exercise of human discernment – merely accepting the first outputs of ChatGPT o1 or Claude 3.5 Sonnet would have led to a decidedly sub-par experience. Teaching students effective discernment and prompting as part of their education is another desirable outcome of this process.

For instance, at Macquarie University, academics in the Faculty of Arts were exploring how to integrate LLMs into their assessments. Across multiple units, from the dedicated Techniques of AI through to German Studies, Ballsun-Stanton has started to observe useful engagement in student responses using the "editor-of-LLM output pattern." The most effective students were those who engaged with LLMs in the "trust, but verify" stance – never taking any of its factual assertions at face value, but using them as a framework to search upon. A student reflected on this when they said: "However, it's crucial to note that AI tools also showed limitations, such as generating false information in legal contexts. This highlights the importance of critical evaluation and fact-checking when using AI-generated content." Here, the best students were critical editors, both of their own and each other's prompts:

> Throughout this unit, I've engaged in various peer mentoring activities that have significantly impacted my colleagues across streams. One of the most valuable experiences has been the collaborative exploration of effective prompting techniques. By examining my peers' prompts, I gained insights into how we all applied Brian's advice for effective prompting. We learned to pay attention to punctuation, populate the context window appropriately, provide specific roles when relevant, and prioritise one task per point in a thread. However, I noticed that many of us, myself included, often struggled to consistently implement all these elements when initiating a new prompt.
>
> This observation led me to appreciate the value of the "prompt grimoire" concept. Having a collection of well-crafted prompts at our disposal proved immensely helpful. It allowed us to build upon each other's work, saving time and effort in constructing effective prompts from scratch. I found that I could easily pick up a peer's prompt that had already gone through the "effective prompt checklist" and adapt it to my specific needs.

In these units, students with the instructor's explicit support and permission, were sharing not only their outputs, but the user and system prompts and their own evaluation techniques, with each other. They learned how to be effective and critical consumers of LLM outputs, figuring out *patterns* for when the LLM output would be useful to them. To train this skill, an early assessment had students critiquing their own prompt logs, looking for when their prompts produced effective or ineffective output and asking students to link specific passages in their prompts to the consequent effective or problematic outputs.

In our worked example, Claude raised the objection: "|W3| There's an unexamined assumption that temptation is necessary for developing executive virtues." Notably, this objection was not in the material from our original essay, which therefore would allow a student to reflect on this old theological argument around class-based distinctions for salvation. While, to us as philosophers, this series of arguments is well tread ground – to a student in an introductory philosophy unit, these arguments for and against virtue ethics and its consequences could be genuinely novel.

Furthermore, as philosophers of understanding tend to agree, understanding is a particular type of knowledge: knowledge of causal and conceptual interrelations (Grimm, 2021). Thinking through multiple potential objections to a conclusion and their interrelations is thus potentially a way to acquire understanding. After all, one's response to the first objection could be inconsistent with one's response to the second objection. Or one might see that all of the objections rely on the same flawed (or true) premise. And so on. Working through scaffolded proleptic reasoning using an LLM is thus potentially a way to abandon a mistaken conclusion, to acquire knowledge rather than mere true belief in a correct conclusion, and — most ambitiously — to acquire understanding of the debate surrounding one's (in)correct conclusion. These are not

inconsiderable epistemic achievements. Effectively, this "conversation" with theoretical objections can elevate an engaged student's argumentation from an "analyze/apply" level in Bloom's Taxonomy to a thoroughly evaluated argument (Armstrong, 2010).

One example here is Kudina's integration of LLMs in philosophy education from 2022 onward as both an object of reflection and a digital skills method that hints at their effectiveness in fostering proleptic reasoning skills in students, while allowing them to acquire knowledge and experience in the workings of LLMs (Ceres, 2023). In her graduate and undergraduate classes at TU Delft and Yale University, she asked the students to use LLMs to prepare debate positions on the desirability of AI use in socially controversial cases, assigning them several topical readings for homework ahead of class. A first sample prompt, upon which the students were invited to build, was the following: "Prepare two arguments and three counterarguments to defend a position in a philosophy debate, titled 'We should use AI for decision-support in matters of social benefits allocation.' Use up to 250 words, justify with academic references." In reworking and responding to the prompts, surveying the LLM-suggested sources and proposed arguments and counterarguments, the students prepared substantiated debate positions and could anticipate the counter-arguments of their classmates. Additionally, they acquired first-hand experience in the misinformation risks of LLMs, spotting the made-up academic references in many instances. The car metaphor is instructive here, as the students need to first know how to properly reference academic sources themselves before challenging the LLM-generated reference list, and similarly, process academic readings to identify factual and logical inaccuracies in LLM-suggested arguments. Kudina also pointed out to the inclusion benefit of using LLMs in class, as it allowed her non-native English-speaking students to formulate and voice their thoughts faster than without the use of LLMs (Ibid.). Even though such examples are non-conclusive, they point to a potential of using LLMs in class and a need to formally evaluate their effectiveness (Cotton et al., 2023).

Whether a student manages to master the skills of crafting an evaluated argument will depend on their background knowledge, their discernment, their epistemic motivation, the effectiveness of their prompts, the capabilities of the specific LLM in question, and the quality of the output from the LLM. There is no guarantee that this will work, but then there are almost never guarantees in teaching and learning. The trick is in the meta-reflection on ChatGPT's output. By requiring multiple modes of engagement, we scaffold more opportunities to engage with the source material and the student's own beliefs across of all levels of Bloom's taxonomy (Armstrong 2010). We envision that students might also harness this experience, however, to scaffold genuinely social proleptic reasoning among themselves. In our experience in the classroom, students often find it difficult or unkind to challenge each other's ideas. To the extent that they learn from proleptic scaffolding via an LLM, they may become more willing and inclined to challenge each other in civil and charitable ways, and to respond to challenges with good faith engagement rather than hurt feelings. Ballsun-Stanton has observed multiple students exploring private arguments with an LLM, treating it as a consequence-free zone to explore their ideas without "looking bad" in front of their classmates. There was also the observation that, as they were engaging in a wargame using Claude to support their arguments, they were able to distance themselves from the claims they were making, and more inclined to view counter-arguments not as attacks on themselves, but on the character the LLM was helping them to roleplay. Thus, we envisage that LLMs could be used not only as solo technological scaffolds but also as a way to build towards genuinely socially scaffolded and collaborative inquiry. Early

anecdotal evidence suggests that this may already be the case with students who are on the autism spectrum (Hoover and Spengler, 2023).

If this is right, then LLMs may also contribute to the sort of public discourse that John Stuart and Harriet Taylor Mill (1859/1989) recommended in *On Liberty*:

> However unwillingly a person who has a strong opinion may admit the possibility that his opinion may be false, he ought to be moved by the consideration that, however true it may be, if it is not fully, frequently, and fearlessly discussed, it will be held as a dead dogma, not a living truth. [...] Whatever people believe, on subjects on which it is of the first importance to believe rightly, they ought to be able to defend against at least the common objections.

Here the calculator metaphor is especially apt. The Mills emphasize that, in order to avoid holding onto mere dead dogmas, people need to "be able to defend against at least the common objections." Because the current generation of LLMs chooses between a set of likely next-words in a sequence, as predicted by the coincidence of words in their training set (Wolfram 2023), they are especially suitable to generating "the common objections." This feature is associated with undesirable outcomes in other cases, e.g., in the perpetuation of common stereotypes (Li & Bamman, 2021; Abid et al., 2021). However, in the context of civic debate, knowing what the common stereotypes are and how best to argue against them is valuable. If this is right, then LLMs may help prepare students to engage civilly with compatriots and others who endorse such stereotypes and the conclusions that people tend to draw from them. Because these tools are arguably not capable of original or inductive analysis – this technique effectively samples from common arguments on the internet, and applies these common argumentative patterns to the prompt in question. Specifically, insofar as an LLM is predicting the most likely next word based on the statistical patterns of words in its training corpus (Wolfram 2023), Ballsun-Stanton has found that these tools are highly effective at deductive transformations, but lack the necessary judgement for inductive operations with present deployments, though this may be a matter of LLM scale. Thus, just as a student may become accustomed to using a calculator for common operations and thereby develop a mental heuristic to detect errors from miskeyed inputs, we believe that students may develop the same sense for arguments and prompts. Or, at the very least, remember the scaffolded editing and argumentation steps prompted by the LLM, and think to apply them to their other experiences. A student in Ballsun-Stanton's class reflected on the start of this instinct by saying: "This use of AI as a learning aid shows its potential in enhancing critical thinking, which is a skill that transcends disciplines. However, these benefits only become apparent when students know how to ask the right questions and understand the AI's limitations. Teaching students how to refine their prompts iteratively is essential. We found that learning to craft effective prompts was key to getting the most out of [the AI]." However, in a dangerous turn, many of the low-effort assessments turned in by students in a colleague's critical thinking class all had the exact same argumentation structure (and flaws) due to the free version of ChatGPT responding in virtually identical fashion to effectively the same prompt. These tools *cannot* work well without the effective application of human judgement to their output.

Finally and relatedly, we believe that the proleptic use of LLMs may help students to recognize and counter confirmation bias, which is a pervasive problem even among scientific experts. Consider the following "golden rule" that Charles Darwin (1887/2009) formulated for himself:

> whenever a published fact, a new observation or thought came to me, which was opposed to my general results, to make a memorandum of it without fail and at

> once; for I had found by experience that such facts and thoughts were far more apt to escape from the memory than favorable ones.

Darwin was hardly a slouch when it came to scientific inquiry, but even he found that he needed to use heuristics and tricks to help himself overcome confirmation bias. But what is confirmation bias? For the purposes of this paper, we conceptualize confirmation bias as a suite of psychological dispositions related to the questions we ask, the evidence we take seriously, the way we interpret evidence, the interlocutors we talk to and how we talk to them, and so on (Nickerson, 1998). At multiple stages of inquiry, people are disposed to engage in these processes in ways that lead them to hold on to their cherished prior beliefs and avoid or discount evidence that these beliefs may be epistemically flawed. As Mercier & Sperber (2017, see also Alfano, 2019) have shown, in cases of solitary reasoning, confirmation bias tends to run riot. However, in some contexts of group reasoning — especially adversarial but civil group reasoning — the confirmation bias of one side tends to cancel out the confirmation bias of the other side, and together these biases ensure that a wider range of evidence and reasons are taken into account than would be otherwise. In such cases, individual confirmation bias is not eliminated. Instead, it is harnessed to support better collective decision making and inquiry.

Our contention is that LLMs can be used to scaffold the adversarial group reasoning that Mercier and Sperber valorize. This point is related to but distinct from our discussion of the *Meno* above. For Plato, what makes certain kinds of knowledge more valuable than true belief is the tether of reason, and we suggested that one way to acquire this tether is through proleptic dialogue with an LLM that is prompted to generate counterarguments and counterexamples. Our point here is that human reasoning itself operates differently in social versus solitary contexts. And in certain types of adversarial contexts, whether a tether is provided or not, confirmation bias is harnessed rather than eliminated in a way that leads to better reasoning by the *group*, even if the *individuals* in the group do not reason any better. By encouraging students to challenge their own or classmates' arguments, rather than to engage in Carnapian confirmation, we allow poor arguments to be defeated by common challenges and stronger arguments to develop effective prebuttals – simulating an evening's debate and discussion amongst friends.

Such group reasoning, we suggest, can be scaffolded using an LLM with the right kind of prompting, which students can be taught to do. Students in an LLM-supported wargame, roleplaying as the United States Cabinet debating 1947 policy, commonly entered the opposition's arguments along with their own plans to help figure out their own responses. And as before, we think that this technological scaffolding could in some pedagogical contexts be the first step to genuinely social inquiry with other humans. If students learn to spar with non-sentient, non-judgmental ChatGPT in a way that they can recognize makes for better reasoning, they may become more willing to do the same with their classmates. Even if this approach does not mitigate the social anxiety of telling a peer that "that argument might be incorrect," the use of these tools offers a judgment-free experience with an endlessly patient, if drunk, tutor willing to push upon and debate the student's ideas. While this experience is not a pure substitute for an adversarial debate between two thinking people, it can provide an adequate simulacrum of such.

Finally, we want to address the relationship between LLMs and diversity, especially the imperative for students to engage seriously with scholarship by researchers from a diverse range of backgrounds on multiple dimensions, including but not limited to gender, race, ethnicity, nationality, and class. This issue is especially pertinent in the discipline of philosophy, where just

a handful of "genius" white men receive the lion's share of attention and citations,[5] but it is no doubt a problem in many disciplines. For both epistemic reasons and moral/political reasons, this disproportionate focus is lamentable or even deplorable. As a discipline, we will learn more and engage with a wider range of considerations and arguments if we are more inclusive. In addition, members of minoritized groups may fairly complain of epistemic injustice when their epistemic contributions are systematically not engaged with, recognized, and respected (Fricker, 2007). In particular, receiving zero uptake in the form of citations to serious scholarship, while others receive more than their fair share of uptake could arguably qualify as a kind of silencing (Dotson, 2011).

While LLMs often embed various pernicious stereotypes, they are often (though not always) able to associate published scholars with various demographic characteristics. Using the service Perplexity.ai, offering a LLM interface (Claude 3.5 Sonnet) to a search-engine, we first gave it the following prompt: "Please act as my research assistant. Here we are going to be looking up notable philosophers of different disciplines and topics. First, can you please give me a list of philosophers who are notable for working on the 'ontological argument? ' Please also include a brief summary of their work."[6] It returned St. Anselm of Canterbury, René Descartes, Gottfried Wilhelm Leibniz, and Immanuel Kant --- all with citations.

However, when prompted to return non-Christian philosophers who addressed the ontological argument, it returned Avicenna, Maimonides, and Al-Ghazali. All men, none Christian. Next, when prompted to return women philosophers who addressed the ontological argument, it returned a somewhat plausible list (Mary Astell and Iris Murdoch). While this may not be the most promising initial list, if a student were engaging with a chat-based LLM with the "drunk tutor" metaphor in mind, it might at least point them in the direction of sources that they could and should engage with *so long as they thought to ask the correct question*. Google Scholar is incapable of offering these sorts of responses, and in many publications the author's name is shortened to a single initial, making it impossible to guess gender. We think that this example suggests that thoughtful and critical use of LLMs could help students to develop more diverse and inclusive reference lists that would both benefit their work from an epistemic point of view and potentially reduce the prevalence of silencing.

Discussion

In this paper, we hope to have moved beyond the hype and doomsaying surrounding chat-based LLMs in the education sector to a more sober and detailed account of *some potential* pedagogical uses of LLMs. In particular, we argued that the different metaphors — a calculator, a car, a drunk tutor — available in the zeitgeist suggest different functions and affordances for a student's relation to LLMs. Each metaphor sheds some light on the phenomenon while obscuring other aspects, as metaphors are wont to do. We find the drunk tutor metaphor especially helpful for thinking about the use of LLMs in the scaffolding of proleptic reasoning, i.e., the anticipation, charitable articulation, and response to potential objections in the form of counterexamples, counterarguments, and so on. LLMs sometimes, indeed often, produce

---

[5] See, for instance, this blog post by sociologist Kieran Healy on philosophical engagement with David Lewis versus all of the women in philosophy (url = < https://kieranhealy.org/blog/archives/2013/06/19/lewis-and-the-women/ >, accessed 26 August 2023).

[6] For the full conversation log, see https://osf.io/yv6ra

adequate or even exemplary text in response to various prompts. But they are stochastic and unreliable. This means that their outputs should generally not be totally ignored but also that they cannot be taken at face value or uncritically.

Additionally, when considering how to integrate LLMs in class, educators need to do so with specific learning goals in mind, carefully weighing the potential benefits against the downsides, related but not limited to the privacy concerns (Kim et al., 2023), large energy costs of training (Strubell et al., 2019) and water costs of using (Li et al., 2023) LLMs, going along with using the tool developed with ethically-problematic practices (e.g. the reinforcement learning for ChatGPT done by underpaid workers in Kenya [Perrigo, 2023]) and contributing to further entrenching the corporate role in structuring academic practices (e.g. GPT models embedded in the Microsoft Office ecosystem [Murphy Kelly, 2023]).

While we do not turn a blind eye to the sociopolitical problematics of LLMs, in this paper, we aimed to suggest that their use also offers a considerable potential to the advancement of learning. As it turns out, some of the technical flaws of LLMs can be recast as features in the context of scaffolding proleptic reasoning. When the outputs are high-quality, verifying that they are high-quality contributes to learning. When the outputs are wrong, likewise, verifying that they are wrong contributes to learning. In all cases, the critical mode required for effective use of a Large Language Model increases the necessary student engagement and promotes critique and argument over mere rote rephrasing. Thus, LLMs can help students acquire and, more importantly *test*, the tether that Plato suggests distinguishes knowledge from mere true belief and makes knowledge more valuable than mere true belief. At the same time, learning to spar with an LLM may help students become better at sparring civilly with the ideas of fellow citizens. And learning to simulate adversarial inquiry with an LLM may help them to become better at engaging in adversarial inquiry with other students and, after university, co-workers and fellow citizens. Finally, as part of a practice responding to the tendency to engage in silencing and other forms of epistemic injustice, LLMs may help train students to attend thoughtfully and consciously to what would otherwise be blind spots in their research process. For these reasons, we think that there is potential for the productive use of LLMs in the classroom. This is not to suggest that worries about academic misconduct using LLMs are overblown, but we do suggest that there are positive uses — especially in the scaffolding of proleptic reasoning — alongside the negative ones.

Additional Information

Ethics. Student Reponses were gathered with specific informed consent under Macquarie University HREC Project 16084. Students participated in three special streams of ARTS3500, the faculty capstone unit, which were dedicated to domain-specific investigations (Ancient History, Politics and International Relations, and Philosophy) within the context of Techniques of AI.

Data Availability. Deidentified student responses and data are available under mediated access. Students have consented to the following clause: “Your participation is completely voluntary, and your decision to participate or not will not affect your grades. All the information collected, including transcripts from your use of LLMs, assessments, and optional surveys and interviews, will be deidentified, meaning your name and identifying personal details will be removed. The deidentified data will be securely stored and may be shared with other researchers in the future

doing similar work with appropriate Ethics approval for re-analysis. The findings from this study will be published in articles and presentations, but your identity will never be revealed. If you have any questions or concerns, you can contact the researchers at any time." As data cleansing is still underway at time of submission, interested researchers are encouraged to contact brian.ballsun-stanton@mq.edu.au for access to the data. It will eventually be published under mediated access on the Australian Data Archive.

Worked examples from the authors are available at: https://osf.io/kn2aq/

References

Adams, F., & Aizawa, K. (2001). The bounds of cognition. *Philosophical psychology*, *14*(1), 43-64.

Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, *3*(6), 461-463.

Alfano, M. (2019). Nietzsche's affective perspectivism as a philosophical methodology. In P. Loeb & M. Meyer (eds.), *Nietzsche's Metaphilosophy*. Cambridge University Press.

Armstrong, P. (2010). "Bloom's Taxonomy." Vanderbilt University Center for Teaching. Accessed on September 27, 2023 at https://cft.vanderbilt.edu/guides-sub-pages/blooms-taxonomy/.

Bonger, S., van der Wal, R., & van der Veldt, M. (2023, January). *TU Delft teachers on ChatGPT: "Banning it is pointless."* TU Delta. Accessed on September 27, 2023 at https://www.delta.tudelft.nl/article/tu-delft-teachers-chatgpt-banning-it-pointless

Ceres, P. (2023). "ChatGPT Is coming for classrooms. Don't panic." *Wired*, January 26. Accessed on September 15, 2023 at https://www.wired.com/story/chatgpt-is-coming-for-classrooms-dont-panic.

Ceres, P. and Hoover, A. (2023). "Kids Are Going Back to School. So Is ChatGPT." *Wired*, August 23. Accessed on September 15, 2023 at https://www.wired.com/story/chatgpt-schools-plagiarism-lesson-plans/.

Clark, A. (1997). *Being There*. Cambridge, MIT Press.

Clark, A. (2002). Towards a science of the bio-technological mind. *International Journal of Cognition and Technology*, *1*(1), 21-33.

Clauss, Patrick (2007). Prolepsis: Dealing with Multiple Viewpoints in Argument. In Christopher W. Tindale Hans V. Hansen (ed.), *Dissensus and the Search for Common Ground*. Ossa. pp. 1-17.

Cooper, J. M., & Hutchinson, D. S. (Eds.). (1997). *Plato: complete works*. Hackett Publishing.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1-12, https://doi.org/10.1080/14703297.2023.2190148.

Darwin, C. (1887/2009). *The Autobiography of Charles Darwin: 1809-1882*. Classic Books.

Dotson, K. (2011). Tracking epistemic violence, tracking practices of silencing. *Hypatia*, *26*(2), 236-257.

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning and Teaching*, *6*(1).

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology1. *Behavioral and brain sciences*, *3*(1), 63-73.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Fuchs, K. (2023, May). Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse?. In *Frontiers in Education* (Vol. 8, p. 1166682). Frontiers.

Greco, J. (2009). The value problem. in Haddock, Millar & Pritchard, *Epistemic Value*, 313–321. Oxford University Press.

Grimm, S. (2021). Understanding, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2021/entries/understanding/>.

Hoover, A. and Spengler, S. (2023). "For Some Autistic People, ChatGPT Is a Lifeline." Wired, May 30. Accessed on September 19, 2023 at https://www.wired.com/story/for-some-autistic-people-chatgpt-is-a-lifeline/.

Howell, C.W. (2023). "Don't Want Students to Rely on ChatGPT? Have Them Use It." *Wired*, June 6. Accessed on September 15, 2023 at https://www.wired.com/story/dont-want-students-to-rely-on-chatgpt-have-them-use-it/.

Johnson, A. (2023). "ChatGPT In Schools: Here's Where It's Banned—And How It Could Potentially Help Students." *Forbes*, January 31. Accessed on September 15, 2023 at https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/?sh=6d23736b6e2c

Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., & Oh, S. J. (2023). Propile: Probing privacy leakage in large language models. *arXiv preprint arXiv:2307.01881*.

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv preprint arXiv:2304.03271*.

Lucy, L., & Bamman, D. (2021, June). Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding* (pp. 48-55).

McCallum, S. (2023). "ChatGPT banned in Italy over privacy concerns." *BBC*, April 1. Accessed on September 15, 2023 at https://www.bbc.com/news/technology-65139406.

McClure, T. (2023). "Supermarket AI meal planner app suggests recipe that would create chlorine gas." *The Guardian*, August 10. Accessed on August 24, 2023 at https://www.theguardian.com/world/2023/aug/10/pak-n-save-savey-meal-bot-ai-app-malfunction-recipes.

Morimoto, R. (2023). "ChatGPT is as Safe as the Driver Behind the Wheel of a Car." *LinkedIN*, April 28. Accessed on September 15, 2023 at https://www.linkedin.com/pulse/chatgpt-safe-driver-behind-wheel-car-rand-morimoto/

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Mill, J. S. & Mill, H. T. (1989). *JS Mill: 'On Liberty' and Other Writings*. Cambridge University Press.

Mollick, E. (2022). "How to... use AI to teach some of the hardest skills." *One Useful Thing*, December 13. Accessed on September 15, 2023 at https://www.oneusefulthing.org/p/how-to-use-ai-to-teach-some-of-the.

Mollick, E. (2023, May 20). *On-boarding your AI intern*. One Useful Thing. Accessed on September 27, 2023 at https://www.oneusefulthing.org/p/on-boarding-your-ai-intern

Murphy Kelly, S. (2023). "Microsoft is bringing ChatGPT technology to Word, Excel and Outlook." *CNN*, March 16. Accessed on September 19, 2023 at https://edition.cnn.com/2023/03/16/tech/openai-gpt-microsoft-365/index.html.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175-220.

Perrigo, B. (2023). "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic." *Time*, January 18. Accessed on September 19, 2023 at https://time.com/6247678/openai-chatgpt-kenya-workers/.

Pritchard, D., Turri, J. & Carter, J. A. (2022). The value of knowledge. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2022/entries/knowledge-value/>.

Shinde, A. (2023). "Taking Control of Language Models with Microsoft's Guidance Library." *Medium*, August 6. Accessed on September 15, 2023 at

https://medium.com/@akshayshinde/taking-control-of-language-models-with-microsofts-guidance-library-e711cd81654b

Sterelny, K. (2010). Minds: extended or scaffolded?. *Phenomenology and the Cognitive Sciences*, *9*(4), 465-481.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

TU Delft. (2023). "AI chatbots in unsupervised assessment." *TU Delft Teaching Support*, June 13. Accessed on September 15, 2023 at https://www.tudelft.nl/teaching-support/didactics/assess/guidelines/ai-chatbots-in-unsupervised-assessment.

Varga, S. (2019). *Scaffolded minds: Integration and disintegration*. MIT Press.

Willison, S. (2023, April). *Think of language models like ChatGPT as a "calculator for words."* Accessed on September 27, 2023 at https://simonwillison.net/2023/Apr/2/calculator-for-words/

Willison, S. (2024, May). *Slop is the new name for unwanted AI-generated content*. Accessed on January 16, 2025 at https://simonwillison.net/2024/May/8/slop/

Wolfram, S. (2023). What Is ChatGPT Doing … and Why Does It Work? *Stephen Wolfram Writings*. Accessed on January 16, 2025 https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, *17*(2), 89-100.

Yu, H. (2023). Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, *14*, 1181712.