Generative AI for Critical Analysis: Practical Tools, Cognitive Offloading and Human Agency

Simon Buckingham Shum

Connected Intelligence Centre, University of Technology Sydney, NSW, Australia

ABSTRACT

Generative artificial intelligence (GenAI) is now capable of performing tasks that we have considered intellectually demanding. There are justified concerns that this will undermine the agency of both educators and students, if tools are poorly designed, poorly used, or imposed — with consequences for education and the future of work. This short paper contributes practical examples pointing the potential for GenAI to promote critical analysis as part of intellectually demanding tasks, by both students and educators. However, this depends on appropriate usage. The paper then briefly discusses how we may balance the benefits and risks of human cognitive offloading to AI, as a perspective on human agency.

Keywords

Generative AI, critical thinking, agency, cognitive offloading

1. Introduction

There are diverse intelligences and dispositions that we need to cultivate in citizens and students, to equip them for the challenges now confronting society [1-3]. One that persists in all lists of '21st century skills' is critical thinking/analysis. One general form that this takes in formal education is the capacity to understand, critique and formulate arguments, which transfers into knowledge work in the workplace. In this paper, I describe how GenAI apps offer new capability in this regard, reporting on tests I have conducted as a continuation of several decades' research into argument visualisation [4-6]. I then describe a second form of critical analysis, namely, distilling a body of ideas into a more succinct summary, the example being the clear articulation of university course learning outcomes. These examples help to demonstrate what AI can now do, which until recently we considered the preserve of humans. This invites a discussion of how we balance the benefits and risks of human cognitive offloading to AI, as a perspective on the broader question of human agency in future human/AI systems.

2. Critical thinking through argument analysis

2.1. Example 1: Identifying the implicit premises in an argument

Arguments are being made constantly in everyday public discourse, as well as within academia. We aspire for citizens to be able to make robust arguments, as well as critique them appropriately. In the philosophy of argumentation, the recurring types of argument have been taxonomized into a robust set of "Argumentation Schemes" [7]. OpenAI's GPT-4 is a sophisticated aid to analysing everyday arguments, as illustrated in Figure 1.

Figure 1: OpenAI's GPT-4 as an argument analyst. (Left) Analyse the implicit premises in the argument *"We must reduce global warming as soon as possible, because the UN Panel on Climate Change has recommended this."* (Right) It correctly classifies this as an *argument from authority*, whose weight rests on specific premises that can be substantiated/attacked.

Figure 2: An argument by analogy[2] (top) which Bing Chat (GPT-4) can critique (bottom)

---

[2] https://twitter.com/ylecun/status/1659332688786882560

2.2. Example 2: Critiquing an *argument by analogy* on social media

Social media platforms such as Twitter have established themselves as influential channels for public discourse and opinion, although the quality of conversation is of course highly variable with platform and community. In a tweet, a well-known AI researcher argued that "AI doomers", who are proponents of strong AI regulation, would also have called for the banning of pens and pencils. This is an *argument by analogy* [7]. Bing Chat (a version of GPT-4 integrated into the Microsoft Bing search engine) was able to critique this claimed analogy effectively (Figure 2).

2.3. Example 3: Analysing an extended argument to create an argument map

The examples so far have been very short: the arguments have made a single 'move', which GPT could recognise and comment on. Let us now consider a more complex case. In March 2023, a large number of eminent thinkers wrote an open letter calling for a pause in building large language models.[3] Achieving widespread media coverage, this provoked extensive debate, including a letter of rebuttal from another set of academics and industry researchers.[4] This seemed an authentically rich argument to test GenAI.

I asked Bing Chat (now Copilot) to access the letter online and identify the key claim and arguments. It provided a reasonable textual summary, output as a set of bullet points summarising key arguments. However, it is well established that students struggle to critique arguments, and that rendering them visually as an argument map can help them understand the key elements of the argument (this is a form of concept map tuned specifically to show multiple perspectives, and the key features of arguments such as supporting/challenging claims/evidence). I asked it to generate a map, but it could not. However, when asked, it confirmed that it understood Argdown, which is a markdown notation for argument maps. It generated this in a code window, which I pasted into the Argdown web app,[5] resulting in a map (Figure 3).

Figure 3: Argdown code generated by Bing Chat from its analysis of a letter, which the user pastes into the Argdown web app to render an Argument Map

Examination of the argument map reveals to what extent this was a rigorous analysis, but also illustrates 'hallucination in argument mapping' (Figure 4). Hallucinations of two types were found. Firstly, the red underline signals *incorrect classification of a premise* using incorrect, or indeed made-up argument schemes. There is to my knowledge no such argument type as *Argument from responsibility,* or *Argument from precaution. Argument from omission* seems to be a jumbling of *Fallacy of omission* and *Argument from ignorance.*

Secondly, there were hallucinated summaries. This node apparently reads well as a summary, but the authors do not talk about researchers at all.

Asking students to perform critical evaluations of AI-generated argument maps should serve as assurance of learning about the subject matter, but can also provide important insights for them into the limitations of AI, if students are equipped and empowered to see through hallucinations.
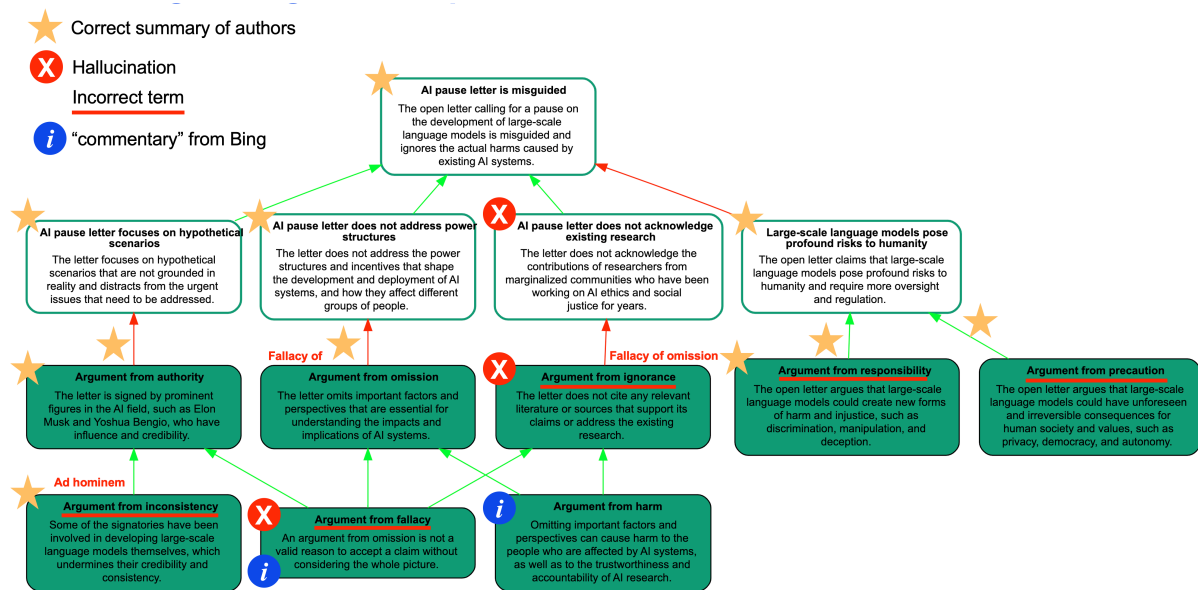

Figure 4: Evaluation of Bing Chat's Argument Map

2.4. Conversing about argument analysis

Conversational agents are exciting for education since they are, by definition, premised on learning through dialogue — hardly a novel concept. But consider this illustration of GPT's capabilities (Figure 5).

Figure 5: An instructive dialogue in which Bing Chat explains very clearly why it added information to the Argument Map that was neither requested nor in the source article

The capacity to add relevant information that was neither requested nor in the source article, and explain this when queried, is unprecedented. In the next example, Bing Chat is asked if can add new nodes to the Argument Map (Figure 6).
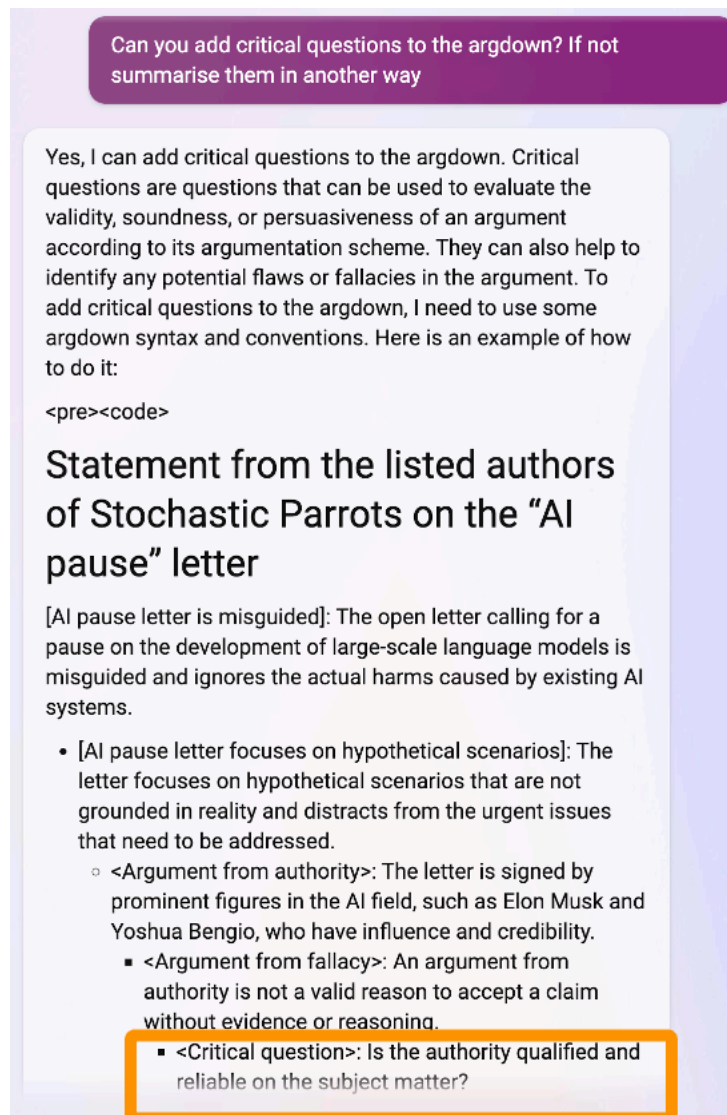

Figure 6: Bing Chat confirms that it can add new nodes to the Argument Map, provided as an instructional device by the academic

Bing Chat's Argdown code can also be rendered as textual outlines. Figure 7 shows the addition of the critical questions, and substitution with placeholders for students to complete.
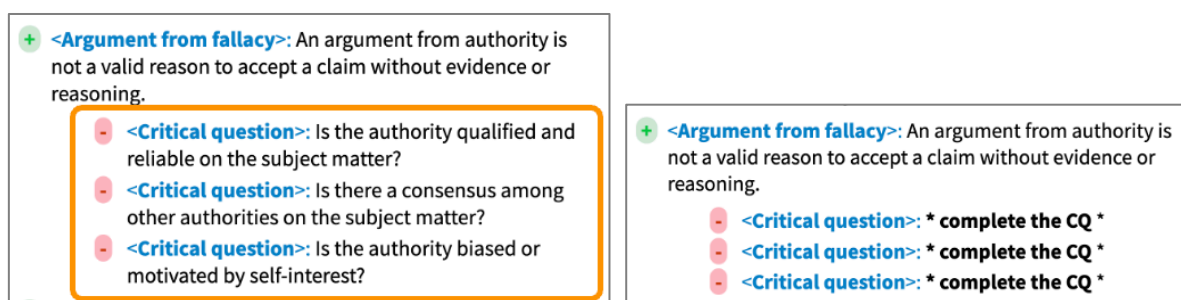

Figure 7: Bing Chat's Argdown code can also be rendered as textual outlines. (Left) Bing Chat's Critical Questions have been added. (Right) Bing Chat is asked to substitute placeholders for students to complete.

It should be noted that the above capabilities are all from the generic ChatGPT-4 model, but as discussed later, customizable intranet GPTs open many new possibilities for tuning chatbots educationally, to institutionally-specific requirements (e.g., within a particular degree program).

An important educational question arises, as we see this kind of performance, namely, will the students engage in excessive cognitive offloading, and fail to learn how to do this themselves? We return to this in the discussion about user agency.

3. CILObot: analysis and summarisation of learning outcomes

Thus far, we have focused on critical thinking and reflection around arguments, primarily with students in mind, but equally, these are tools for any professional to test their thinking. In the next example, we focus on a specifically instructional task, which harnesses the generative capability of LLMs more fully to distill complex text into key themes. The text in this case is a specific 'genre' of writing, the *Course Intended Learning Outcome (CILO)*. CILOs define what students know and can do on successful completion of the course. As part of a well-designed curriculum, each part of a course – subjects, modules and assessments – should all respond to its CILOs. Effective implementation of CILOs requires both the subject matter expertise of academics and the pedagogical knowledge of learning designers (LDs). Indeed, recent evidence points to the benefits that academics gain from working with LDs on their online teaching, and how this transfers to their in-person teaching [8].

One specific element in this task that academics can struggle with is to articulate good LOs. Furthermore, these typically vary widely in quality and quantity between academics. At UTS, we are working towards summarising all courses consistently using approximately six CILOs, to achieve a better user experience as students make enrolment decisions, and to assist teaching teams in their course design and reviews. However, it is an intellectually and linguistically demanding task to distill a list of 20-30 CILOs (which is not uncommon), down to six well designed CILOs, and the university needs to implement this summarization for its entire program.

It is here that we anticipated that LLMs could assist. GenAI intranets now provide universities with authenticated, secure, private services, integrated with other internal services, and tuned to support business processes.[6] In a 2-day hackathon, iterative prompt engineering informed by feedback from academics and learning designers led to the refinement of a system prompt that configured 'CILObot', a ChatGPT to aid in drafting these new CILOs. The system prompt incorporates widely recognised design principles (e.g. open each CILO with a verb from Bloom's Taxonomy), with the addition of internal requirements (e.g., UTS Indigenous-CILOs), and the chatbot is grounded in a corpus of documents about CILO design.

The prototype is showing promise, and after a day's intensive work using the Azure 'Chat Playground' (the ChatGPT design environment), the results for several programs in our Health faculty were validated by disciplinary experts (e.g., Figure 8). CILObot generates a coherent first draft in about 30 seconds, which can of course then be refined through further conversation with it, and edited by the teaching team. We estimate that agreeing on how to distill 20-30 CILOs into 6 would normally be a minimum of 3 hours' meeting between the Course Director and the program's lead academics, which represents an impressive return on investment. Next steps will test CILObot with other degree courses.

Figure 8: *UTS CILObot* (top)*,* a university intranet GPT-4 agent, proposes a way to distill 26 Course Intended Learning Outcomes (CILOs) down to the target of six (in bold). The output explains how it has derived them from the originals.

4. Discussion: cognitive offloading and human agency

These capabilities are, in my view, impressive. If students were to produce argumentative reasoning as presented above, we would surely conclude that they were thinking critically, and had mastered some argumentation principles. Similarly, if an academic proposed a set of six distinctive, well expressed CILOs with complete coverage of the original CILOs, we would regard that as exactly the kind of task senior academics should be capable of. The difference, of course, is that these tasks are performed in under a minute, producing coherent drafts.

We do not need to believe that agents have the same kind of understanding as people to appreciate the value of AI being able to communicate with this fluency and precision in order to provoke critical human reflection. GenAI performs these tasks in seconds, and can iterate its analysis as often as requested. In principle[7], therefore, GenAI can be used to:

- offer students, academics or any other kind of analysts instant, formative feedback on draft arguments, for instance, identifying points that could be potentially attacked;
- analyse a written corpus to give insights into the quantity and quality of argumentation, which could inform both LA researchers, practitioners, and educators;
- analyse a written corpus in order to derive a representative set of summary themes (noting that AI cannot 'read between the lines' as a human qualitative analyst does).

---

[7] These are *in principle* capabilities for GenAI argument analysis and feedback, since this has not yet been tested empirically with students, to the best of my knowledge.

The pivotal question — whether we are envisioning the future of learning among students or professionals in the workplace — is the "allocation of function" between human and machine, to use the original term from ergonomics. Questions of cognitive offloading and human agency now arise, as we consider different scenarios.

If AI improves short-term productivity (e.g., faster syntheses of complex information; more creative ideas; more incisive reasoning), we might anticipate (and indeed we are already seeing in certain professions) that AI apps will embed into professional work practices. Professionals are qualified to 'drive' such intellectual power-tools (in contrast to students their qualifications should enable them to recognise poor AI output); they will welcome cognitive offloading in their busy lives; and if they do not use AI may find they are unable to compete with those who do. We might see this as empowering professionals — and yet we might also see a loss of agency as they are essentially forced to use AI in order to compete. Time will tell if the long-term use of AI leads to the degrading of important human capabilities, just as GPS satellite route navigation has for many young people obviated the need, and hence ability, to navigate via printed maps.

In sharp contrast, for education the story is very different. "Productivity gains" need to be judged by a different yardstick, since while an essay written solely by GenAI in 2 minutes is a "productivity gain" in terms of artifacts/minute, the absence of the student's cognitive engagement fails other "KPIs" for meaningful education. Students must build their foundational knowledge, skills and dispositions, in order to function as citizens and professionals in the myriad contexts in which they cannot call on AI, but must think on their feet and demonstrate diverse intelligences [1, 2].

Consequently, as emphasised in a recent national report for the higher education sector, assessment must be reformed for the age of AI [9]. Cognitive offloading takes on special importance in assessment design [10], since it forces us to ask what exactly we deem important to assess in the age of AI. It is beyond the scope of this short paper to expand on this issue further, but a fitting conclusion is to return to AIED research 30 years ago, and remember a distinction made by Roy Pea (emphasis added):

> *"Pedagogic systems* focus on cognitive self-sufficiency, much like existing educational programs, in contrast to *pragmatic systems*, which allow for precocious intellectual performances of which the child may be incapable without the system's support. We thus need to distinguish between systems in which the child uses tools provided by the computer system to solve problems that he or she cannot solve alone and systems in which the system establishes that the child understands the problem-solving processes thereby achieved. We can call the first kind of system pragmatic and the second pedagogic. Pragmatic systems may have the peripheral consequence of pedagogical effects, that is, they may contribute to understanding but not necessarily. The aim of pedagogic systems is to facilitate, through interaction, the development of the human intelligent system. While there is a grey area in between and some systems may serve both functions, clear cases of each can be defined." [11]

GenAI forces us to ask when we are — or should be, as the boundary shifts — assessing joint human+AI system performance, versus capability without AI. A consequence of this distinction is that we must cultivate "*mindful engagement*", not "*mindless engagement*" [12]. In the intense debates about whether the human (student, academic or professional) remains sufficiently in the loop, these concepts from the era of symbolic AI, when they could barely glimpse what is now possible, remain as important as ever.

Acknowledgements

References

[1] H. Gardner, *Five Minds for the Future.* Harvard Business Review Press, 2009.

[2] D. G. Thompson, "Marks Should Not Be the Focus of Assessment – But How Can Change Be Achieved?," *Journal of Learning Analytics,* vol. 3, no. 2, pp. 193-212, doi: 10.18608/jla.2016.32.9.

[3] L. Markauskaite *et al.*, "Rethinking the entwinement between artificial intelligence and human learning: What capabilities do learners need for a world with AI?," *Computers and Education: Artificial Intelligence,* vol. 3, p. 100056, 2022, doi: https://doi.org/10.1016/j.caeai.2022.100056.

[4] S. Buckingham Shum, "The Roots of Computer-Supported Argument Visualization," in *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*, P. A. Kirschner, S. Buckingham Shum, and C. Carr Eds. London: Springer-Verlag, 2003, pp. 3-24.

[5] S. Buckingham Shum, "Sensemaking on the Pragmatic Web: A Hypermedia Discourse Perspective," presented at the 1st International Conference on the Pragmatic Web, 21-22 Sept 2006, Stuttgart, 2006. [Online]. Available: Open Access Eprint: http://oro.open.ac.uk/6442.

[6] S. Buckingham Shum, "Cohere: Towards Web 2.0 Argumentation.," presented at the 2nd International Conference on Computational Models of Argument, 28-30 May 2008, Toulouse, France, 2008. [Online]. Available: http://oro.open.ac.uk/10421.

[7] D. Walton, C. Reed, and F. Macagno, *Argumentation Schemes*. Cambridge: Cambridge University Press, 2008.

[8] D. A. Joyner, A. Rusch, A. Duncan, J. Wojcik, and D. Popescu, "Teaching at Scale and Back Again: The Impact of Instructors' Participation in At-Scale Education Initiatives on Traditional Instruction," presented at the Proceedings of the Tenth ACM Conference on Learning @ Scale, Copenhagen, Denmark, 2023. [Online]. Available: https://doi.org/10.1145/3573051.3593389.

[9] J. M. Lodge, S. Howard, M. Bearman, P. Dawson, and Associates, "Assessment reform for the age of Artificial Intelligence," Tertiary Education Quality & Standards Agency (TEQSA), Australian Government, Canberra, AUS, 2023. [Online]. Available: https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assessment-reform-age-artificial-intelligence

[10] P. Dawson, "Cognitive Offloading and Assessment," in *Re-imagining University Assessment in a Digital World*, M. Bearman, P. Dawson, R. Ajjawi, J. Tai, and D. Boud Eds. Cham: Springer International Publishing, 2020, pp. 37-48.

[11] R. D. Pea, "Integrating Human and Computer Intelligence," in *Children and Computers: Directions for Child Development (No. 28)*, E. L. Klein Ed. San Francisco: Jossey Bass, 1985.

[12] G. Salomon, D. N. Perkins, and T. Globerson, "Partners in cognition: extending human intelligence with intelligent technologies.," *Educational Researcher,* vol. 20, no. 3, pp. 2-9, 1991, doi: 10.3102/0013189X020003002.