Qeios

Peer Review

Review of: "Rethink Your Mental Model in the Age of Generative AI: A Triadic Framework for Human-AI Collaboration"

Maha Bali[1]

1. Center for Learning and Teaching, American University in Cairo, Egypt

Rethink Your Mental Model in the Age of Generative AI: A Triadic Framework for Human-AI Collaboration

Overall, this was a really helpful paper that simply and clearly outlines some important basics related to how AI works, how humans tend to think about AI, and the kind of metacognitive literacies we need to develop along with AI fluency/literacy to ensure we maintain human agency while recognizing the boundaries of what LLMs can do. I can imagine using it and the procedures it recommends in courses at universities and in trainings in organizations.

I find it useful that you begin with the term "**jagged intelligence**" and bring up why recognizing this complexity and seeming non-binary-ness of AI versus more deterministic models of looking at tech is needed as a lens to think about AI/LLMs. I had heard the term before and, of course, recognize the sensibilities you describe, but now I realize how important it is to take on this lens.

Early on in section 2, you refer to chapter 5 and chapter 8. Unsure if this article I'm reading is meant to be part of a book, rather than a journal article? In any case, it felt inappropriate to mention these here for me as a reviewer who does not have access to these chapters, you know? Or do you just mean sections 5 and 8? **I think you just need to replace every reference to "chapter" with "section".**

Sections 2.1 and 2.2 are well-written and simply argued. Your points are clear, and you are building the argument smoothly.

The section 2.5 on AI outperforming humans on creativity and empathy is less convincing; partially, this is a cognitive bias of mine, or my understanding of creativity and empathy, and my own experiences

 doi.org/10.32388/UFPJLI

interacting with AI, but trying to be more objective about it, I think it depends on how you define and measure these very complex constructs (creativity and empathy), and how the studies likely involve short-term text chats rather than more long-term and authentic scenarios?

Regarding failures of LLMs to do tasks traditionally simple for machines, I usually give the LLM this prompt: Write code to do X, then run that code on Y. That usually fixes the problem because the problem does not require the "generative" capability of the LLM, but can easily be solved with a straightforward algorithm... so then I ask the LLM to generate the algorithm and run it 😀 instead of trying to solve the problem directly.

I wonder if what you're trying to say is that LLMs are not good at "less precise" tasks such as calculations and can be better than some humans or even experts at fuzzier tasks? And in that sense, less precise tasks are really easy to prove/confirm the weakness of LLMs, whereas with fuzzier tasks, judgment is more complex and also humans' judgment of their own capabilities differs, so it depends, really, on who is using and judging the quality of the AI output, right?

Figure 1 is helpful, but there isn't somewhere in the paper where you sum the three up together in text, and the figure does not really add any value visually, so I would recommend just removing the figure and putting the text into the body of the paragraph before introducing the new section.

Part II begins with "we need to". While I don't disagree with your suggestions broadly speaking, I think to strengthen your argument, you might want to: **a) define who "we" are? And b) explain why we need to.** I guess you mean that we need to do these things in order to go forward with AI in a constructive/productive way? That we need a deep and realistic understanding of the limitations and possibilities of AI, how our own mental models get in the way, and therefore become better judges of how to use it for our benefit, not our detriment? Something like that?

In section 3.4, I have strong objections to this "Scholars propose shifting toward human-inspired evaluation schemes that consider natural variance and contextual expectation. This reframing aligns assessment with the stochastic nature of generative intelligence". This is not a very convincing argument. You're basically saying that scholars say that since LLMs don't do so well on our "old" mechanisms for evaluating correctness, we should modify our evaluation criteria? While I understand that the reality is the world is more complex than black and white, we also need to adjust our expectations depending on the context, and whether "softer" evaluation criteria could be harmful in high-risk contexts, or if the softer evaluation criteria might help us take advantage of LLMs (such as in more creative contexts where there is no big harm from AI hallucinating or whatever).

Section 3.5 is interesting and useful in that while generic LLMs don't perform well on some things, other tools built on AI do perform well, so it's important to take the field in its entirety and its development trajectory.

I am no longer in the field of computer science, so I cannot evaluate the technical details shared here, but they seem to be on track and well explained to an academic non-computer scientist audience.

For Table 1 of prompt techniques - it is unclear if these are successful prompt techniques?

For section 4.4, where you discuss augmentation vs. collaboration modes, a useful typology was proposed in this article:

McCarthy, I. P., Hannigan, T., & Spicer, A. (2024). The risks of botshit. Harvard Business Review. Accessible here: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5172718

This sentence is incomplete "Interestingly, even if unpredictable LLMs are able to mitigate some human biases[162][163]." - tie it to the next sentence and remove "on the other side". I guess what you're trying to say is that even if unpredictable LLMs are able to mitigate human bias (I assume inconsistently since unpredictable), they introduce their own biases, right, and that those biases are somewhat more stable and clear than human biases?

Explain what "Cialdini-style prompts" means. I've never heard that term before and had to look it up. Also, should the term "unwanted" be used at all? Do LLMs want and not want? Or do you mean "tasks that LLMs refuse to do at first"? Focusing on the observable action, rather than the intent? Or at the very least, use the same character you use before ¬reasoning, so it would be ¬unwanted.

For section 5.2, I suggest two articles I co-authored that tackle the different dimensions of metaphors used for AI and their harms and benefits and usefulness in promoting critical AI literacy:

- Gupta, A., Atef, Y., Mills, A., & Bali, M. (2024). Assistant, parrot, or colonizing loudspeaker? ChatGPT metaphors for developing critical AI literacies. Open Praxis, 16(1), 37-53.
- Atef, Y., & Bali, M. (2025). Whose metaphor is it anyway? Analysing AI metaphors from positionality and values of speaker and recipient. Critical Studies in Teaching and Learning, 13(2), 128-155.

Section 5.5 is very useful, the part about system 0 and the metacognitive burden; this also explains why it is much harder for younger people whose critical thinking is not yet fully developed to resist AI use - they are following their natural tendency for cognitive offloading, especially for tasks they perceive as not valuable or relevant or interesting.

Table 3 is useful, but some elements of it are unclear, because they're building on things in different parts of the text. Some of them I understood right away, but not all of them, and I'm unsure why some seem clearer than others. **Or should Table 3 appear in part III after you explain each proposition separately?**

Section 7 is useful and clear, with the final figure quite helpful in showing how your work all builds on previous points you made.

Your recognition of the limitations of your work is quite comprehensive.

Regarding your notational intervention, the problem is I'm not sure where to find it on the keyboard if I wanted to re-use it. Perhaps it needed to be some other more commonly found symbol? I get that more common symbols have other purposes/connotations.

I did feel towards the end that the term "routine" implies some kind of stability, whereas we're talking about unpredictable models. I know you're talking about these routines as processes to prevent cognitive offloading and overreliance and trust in AI models, a stable set of steps/processes to help humans remain skeptical, but perhaps a different name than "routines" would better reflect how dynamic the processes would actually be in practice?

Regarding transparency when working within teams/organizations, I would recommend what I do in my classes, which is to ask students to share not just the prompt but also the actual link to their conversation with AI (since it's not usually just one prompt, right?).

I agree with this statement in your conclusion: "Just because AI is available does not mean we always have to use it – sometimes a task is not suitable for AI, and sometimes we consciously choose to do tasks ourselves to learn something new."

Overall, a really useful and well-thought-out article. Thank you for writing it.

Declarations

Potential competing interests: No potential competing interests to declare.