Reflections on using LLMs to read a paper[1]

Lovkush Agarwal
Independent

With
Jacques Thibodeau & Apart Research

Abstract

My initial aim was to create a prototype tool that helps a researcher read an AI safety paper. However, it ended up being some basic experimentation with Claude UI and reflection on how to accelerate research. The main learning is that I - and likely the majority of AI safety researchers - are drastically under-utilizing existing technologies, but it is not clear what is a good way to keep up.

Keywords: Research tooling, AI safety

1. Tool overview

The tool is an AI paper reading assistant. This tool assumes the user has already decided to read a given paper. There is a separate (and likely more important) question of how the user decides which papers to read, which I do not focus on here.

Various possible features / requirements for the tool are:

- Frictionless. E.g. the tool is part of existing pdf readers, like Google Scholar PDF Reader chrome extension or Zotero, or, is somehow part of a suite of acceleration tools in VSCode.
- Summarization. Tool can summarize the relevant parts of the paper, clearly and accurately.
- Diagrams. Create excellent visual aids or diagrams that explain the research, not just explaining in words.

[1] Research conducted at the Research Augmentation Hackathon, 2024

- Access to both text and figures. When you upload a pdf to the Claude chat UI, it only accesses the text.
- Jargon / maths explainer. User friendly mechanism to explain jargon or maths. Somehow personalized to the user's background.
- Pedagogy. Help the user actually learn and remember the paper, as appropriate. Basic function is that the tool asks the user multiple choice questions, but can imagine the tool using cutting-edge pedagogical practice (which is obviously the creation of tons of Anki flashcards. 😛)
- Connections with user's existing knowledge or research ideas. Either automatic (which means somehow getting access to user's knowledge or ideas, e.g. via note taking app, or github repo, or research brain dump, or...) or by guiding the user to find connections themselves. This is related to the pedagogy point above.
- Connections with other papers. Tool can compare with other papers (e.g. by appropriately bringing in other papers into its context), describing differences in findings, methodologies, development of ideas,...
- Ideation. Help create research ideas based on this paper. Either extending the work done in paper, or, combining with other ideas, or combining with user's existing research, or ...
- Code generation. Ability to generate code to reproduce (parts of) the research.

2. Why this tool should exist

Staying up to date with the (relevant) research is an important part of being a researcher. Given the quantity of research being produced, I actually believe that selection of what to read is more of a bottleneck than the reading. I think this is particularly true the more experienced the researcher is, because (presumably) experienced researchers can quickly skim read a paper.

Nevertheless, the reasons this is useful include:

- To reduce barriers to entry for newcomers to the field, in particular for people with non-standard backgrounds.
- Advanced features (e.g. writing code or making connections with researcher's knowledge base) should have large multiplier effects, if done to a good enough quality.

3. Results and next steps

I tested Claude UI on a handful of papers, one I had read before and others that I had not.

- Learned lookahead in Chess AI. (I had read this before)
- Truth is Universal: Robust Detection of Lies in LLMs

- The geometry of truth: emergent linear structure in large language model representations of true/false datasets
- The Remarkable Robustness of LLMs: Stages of Inference?
- Future events as backdoor triggers, investigating temporal vulnerabilities in LLMs

The prompts I ended with - for the summarizing and for testing my understanding - were:

- Summarize the attached paper. Include the following sections in your summary: 1. Main findings and insights. 2. For each main finding or insight, detailed explanation of the evidence and method used to get the finding. 3. Differences with previous research, if applicable.
- Please ask me a series of multiple choice questions about the attached paper to check my understanding. Ask questions to probe my understanding of the key findings and the main evidence to justify the findings. Ask the questions only one at a time making sure I understand. In the multiple choice options, make sure the incorrect options seem sensible so I cannot just guess the correct answer using common sense. Please be brief and terse - you do not need to congratulate me when I am correct or provide assurance if I am incorrect.

My general experience was very positive, and viscerally made me realize just how much productivity I am leaving on the table by not exploring LLM and modern AI tools.

Various reflections on the experience:

- Claude was good at summarizing and going into detail into items when I asked for further explanation.
- The multiple choice questions were OK and had varying quality. Some were trivially straightforward but some did make me have to think and reason a little bit with the ideas.
- The experience forces me to consider what is useful about reading a paper. For the last couple of papers, *I did not look at the original paper*, and my current knowledge of them is based purely on my interactions with Claude. Given my goals as somebody transitioning into AI safety (get broad awareness of ongoing research), I am happy with the level of understanding I gained.
- Big limitation of the UI is you have a limited number of calls with Claude Sonnet 3.5, even with the paid subscription. Using the API should overcome this problem.

Next steps.

This is unclear, but include the following:

- Prioritizing the bottlenecks, rather than going all in on this particular tool.
- Thinking about how to appropriately evaluate the tool(s), especially during the development of the tool and the prompt engineering. I think this is a big

challenge given that most of the things we really care about are difficult to measure.

- Thinking about how to get community actively using beneficial tools. This is challenging because there are so many useful things that each of us could be doing to improve our lives/productivity (exercise, diet, sleep, various habits from rationality, various good habits from software engineering, various habits for research, various LLM habits, etc.), it is hard to know how this is all coordinated and prioritized.