
Tool overview report¹

Luis Tercero Herman

With
Jacques Thibodeau & Apart Research

Abstract

Academic Weapon is a Chrome extension designed to accelerate the onboarding process for new researchers and students in the field of AI Alignment. By leveraging state-of-the-art Large Language Models (LLMs), this tool provides contextual assistance directly within the user's browser, facilitating rapid comprehension of complex AI safety concepts. The extension's key features include a smart side panel for in-situ queries, contextual pre-loading of article content, and quick-access prompts for summarization and analogy generation. These functionalities collectively reduce the cognitive load associated with parsing dense academic literature, allowing users to more efficiently grasp foundational ideas in AI Alignment. By lowering the barrier to entry for newcomers and enhancing the productivity of established researchers, Academic Weapon has the potential to significantly accelerate progress in AI Alignment research.

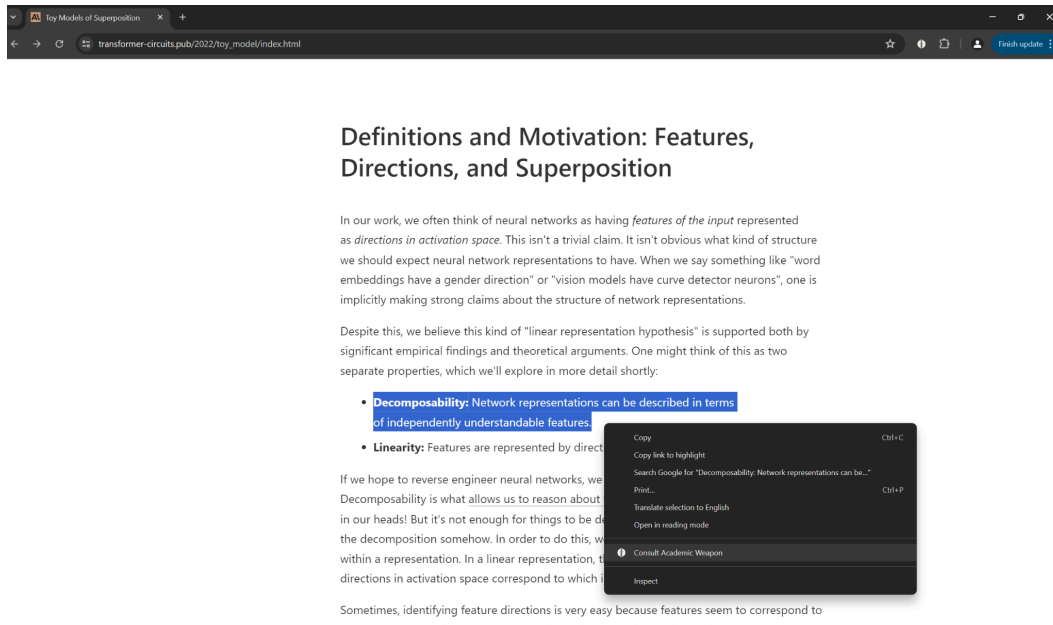
Keywords: Research tooling, AI safety

1. Tool overview

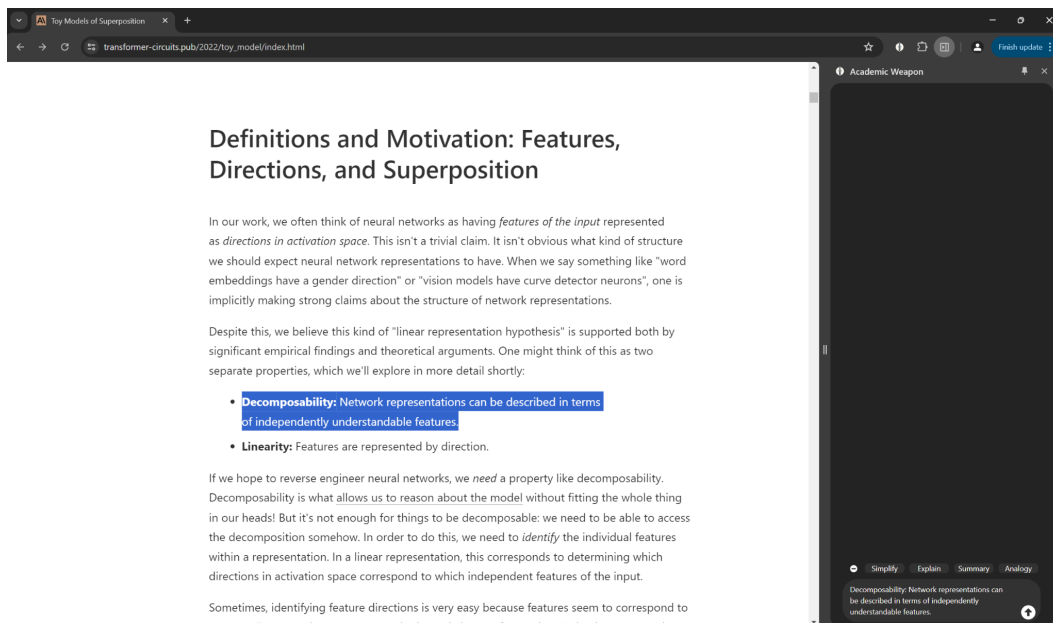
At its core, Academic Weapon aims to address the steep learning curve in AI Alignment research. By making it easier to learn and understand the latest literature, I hope to allow students to contribute to the field much quicker. After installing the [Academic Weapon Chrome extension](#), usage is really simple:

To use it, first navigate to a challenging article (for this example, I will be using [Toy Models of Superposition](#)) and highlight some confusing text. Right-click it, and click “Consult Academic Weapon.”

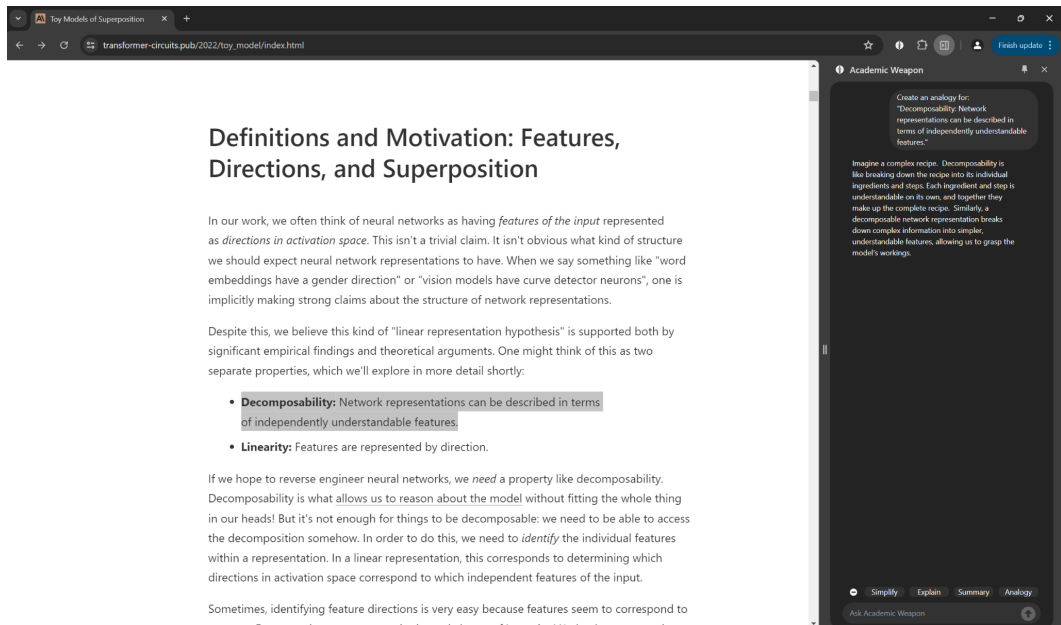
¹ Research conducted at the [Research Augmentation Hackathon](#), 2024



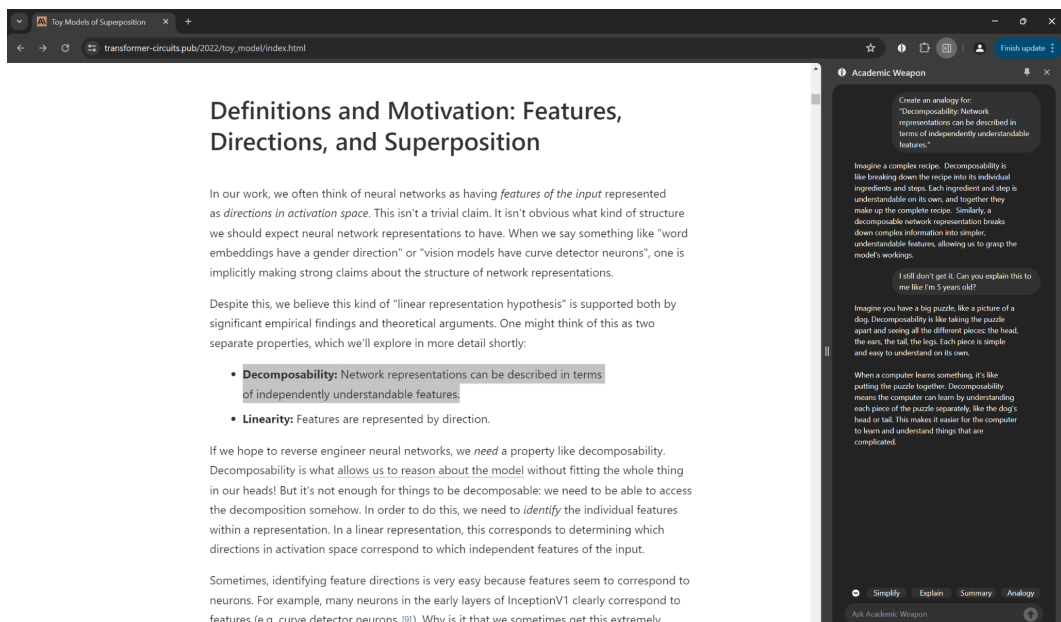
This will open the side panel and copy the selected text into the chatbox.



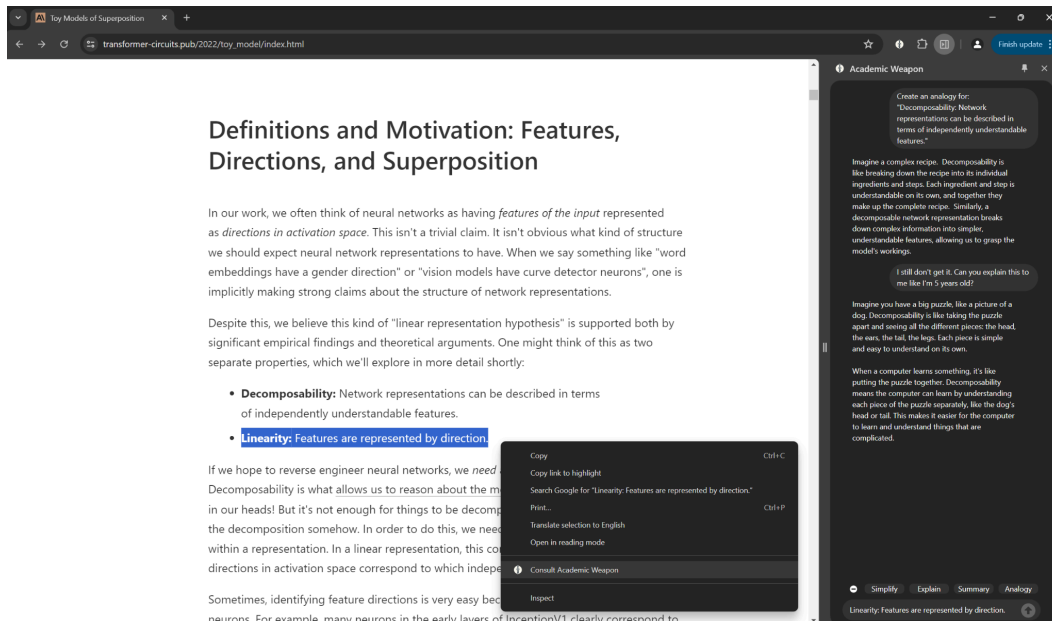
Now, you can ask the LLM any question about this segment of the article! You can either interact with the chatbox and edit the text to ask the LLM a question, or you can click on a Quick Question, a shortcut that allows you to quickly ask the LLM to simplify, explain, summarize, or create an analogy for the selected segment. Here, I asked the LLM to create an analogy for the segment.



But, what if you still don't get it? You can continue to query the chatbot with a custom question any time.



Furthermore, you can continue to right-click to transfer any selected segment to the chatbox!



2. Why this tool should exist

The field of AI Alignment is notoriously complex and rapidly expanding. This tool is primarily designed to lower the barrier to entry for interested researchers. Since it will be easier to understand articles that assume prior knowledge, more people might be incentivized to begin working on AI Alignment problems. Additionally, anyone working on AI Alignment (even experienced researchers) will be able to understand bleeding-edge articles more quickly, which will allow them to contribute to the field more quickly, boosting progress in the field. This is especially important since the urgency of solving AI alignment necessitates rapid progress and efficient knowledge transfer. By reducing the time spent on basic comprehension, the tool frees up cognitive resources for higher-level thinking and problem-solving. This efficiency boost could lead to faster iterations in research cycles and accelerated progress towards ensuring the safety and alignment of advanced AI systems.

3. Results and next steps

The main issue I've found with the extension is its inability to read certain websites. For example, web-hosted PDFs or code from GitHub cannot be accessed by the chatbot. I am currently working on a solution.

4. References

Elhage, et al., "Toy Models of Superposition", Transformer Circuits Thread, 2022.