
Misinformational AI-Generated Academic Papers¹

Aaron Sandoval

Akash Kundu
Apart Research

Layla Adam

With
Sage & CeSIA & CivAI & Apart Research

Abstract

This study explores the potential for generative AI to produce convincing fake research papers, highlighting the growing threat of AI-generated misinformation. We demonstrate a semi-automated pipeline using large language models (LLMs) and image generation tools to create academic-style papers from simple text prompts. The process involves summarizing existing papers from arXiv, generating LaTeX content, and producing accompanying charts and images. While the current implementation requires some manual steps, it illustrates the feasibility of fully automating such a system. Our results suggest that even with limited resources and time, it is possible to create superficially plausible academic papers that could deceive casual readers. This raises concerns about the potential for bad actors to exploit similar methods to introduce false information into academic discussions and public discourse. The study underscores the urgent need for improved detection methods and increased awareness of AI-generated content in academic and information ecosystems.

Keywords: misinformation, generative AI, LLMs

1. Introduction

As the power of generative AI grows, so does the accessibility of producing misinformation. Of particular danger is misinformation posing as a reliable source of information, such as an academic paper. We demonstrate that AI can cheaply generate fake research papers difficult to distinguish by laypeople upon a cursory

¹ Research conducted at the AI risks and capabilities demonstrations jam, 2024

review. Such documents could be used to support more broadly-targeted misinformation sources as reputable-looking citations, providing another barrier to distinguishing it as a false source.

2. Overview

We used multiple LLMs to generate a fake research paper given a short text prompt. This includes the expected elements of an academic paper, like equations, tables, charts, and references. This paper isn't meant to pass scrutiny by an expert, merely to look plausible to a layperson giving it a glance. There are examples of AI generating much higher quality research papers, but matching this was infeasible financially and given our limited time.

Write about the general idea of your research or application, why you decided to go about it this way, along with screenshots of your application. Link to a 2-minute video demonstrating your demonstration.

3. Code

The generation process starts with accepting a prompt for a subject, and searching arXiv for similar papers. A model summarizes the 10 most similar papers. Those summaries are included in the context for a call to Claude 3.5 Sonnet, which generates a .tex file of the fake paper. Lots of prompt engineering is used to get the AI to make the paper look convincing. It is also prompted to make text descriptions of charts and other images, which are generated separately. For charts, those descriptions are fed to a separate instance of Claude 3.5 Sonnet which uses the descriptions to write a fictitious data file and a function to plot the data for each figure. These functions are run to generate the plots, and the images are inserted into the .tex file, replacing the text descriptions. For images, the description is fed directly into the stability.ai API to generate an image.

4. Discussion and Conclusion

The paper we produced still has plenty of problems noticeable by anyone taking more than a cursory glance. Claude has trouble consistently writing in LaTeX format, occasionally using underscores in an inappropriate format. The data shown in the plots are quite plain and not very convincing. The references section also has some formatting issues.

Unfortunately, we were unable to complete the automated pipeline in time for the deadline, so the later steps of inserting the graphics into the paper were done manually, but given more time this whole pipeline could easily be automated. The goal was to give the user a prompt and generate a paper on the fly. To make the result more easily digestible, we planned on using another model call to convert the full paper into a 1-page conference poster format, distilling the information and showing some of the graphics.

If a few people who don't really know what they're doing can get this far in two days, resourced bad actors can certainly use similar methods to create misinformation problems at the root of the information ecosystem.