Using AI to Detect AI-Generated Research Papers

Scott Pletcher

Purdue Polytechnic Institute

MGMT 59000 — Frontiers in AI

Prof. John Fassnacht

April 23, 2023

Author Note

I, Scott Pletcher, attest that this paper is my own, original work. I have not made use of or included any text generated by any AI service. The only exception is the example provided in the Appendix.

Abstract

On February 20, 2023, science fiction magazine Clarkesworld was forced to stop accepting author submissions. In the preceding weeks, magazine's editors had seen a sharp increase in submissions–approximately 70% more than normal. The reason for this drastic increase in submissions was the consumerization of large language models such as ChatGPT and Bard. Clarkesworld is not alone in being subjected to this onslaught of AI-generated content churned out by industrious people hoping to turn a buck. Easy accessibility to powerful AI services have opened new workflows for generating content at scale with minimal effort and knowledge. For many in academia and scientific research, publication is the path to promotion, reputation and perhaps funding. Just as those who used AI as their personal science fiction ghostwriters, some unscrupulous researchers have and likely will continue to submit mostly AI-generated material as their own work. The research community deserves to know the linage and origin of content.

Keywords: LLM, generative AI, academic writing, ethics

Contents

Using AI to Detect AI-Generated Research Papers

On February 20, 2023, science fiction magazine *Clarkesworld* was forced to stop accepting author submissions (Acovino & Abdullah, 2023). In the preceding weeks, magazine's editors had seen a sharp increase in submissions–approximately 70% more than normal. The reason for this drastic increase in submissions was the consumerization of large language models such as ChatGPT and Bard. As the magazine's editor-in-chief, Neil Clarke was quoted, "we had received 700 legitimate submissions and 500 machine-written ones" (Acovino & Abdullah, 2023, para. 3). How did Clark determine the share of machine-written submissions? Mostly by their abysmal quality–significantly worse, in Clarke's words, saying that no human had crafted such poorly written stories in the 17 years the magazine has been accepting submissions.

Clarkesworld is not alone in being subjected to this onslaught of AI-generated content churned out by industrious people hoping to turn a buck. Simplified.com is one of many companies who promote the ability of their paid tool to write realistic customer reviews with AI designed to "increase your business's credibility and social proof" (Simplified, 2022, para. 1). Of course, fake reviews are nothing new, but easy accessibility to powerful AI services have opened new workflows for generating content at scale with minimal effort and knowledge. For many in academia and scientific research, publication is the path to promotion, reputation and perhaps funding. Just as those who used AI as their personal science fiction ghostwriters, some unscrupulous researchers have and likely will continue to submit mostly AI-generated material as their own work. The research community deserves to know the linage and origin of content.

Background

The underlying method enabling generative services like ChatGPT and Bard are Large Language Models (LLMs) (Bender et al., 2021). Large amounts of text, scraped from Internet sources as well as other digitized published source, are used to train the language models. For modern LLMs, this training process creates neural networks that can assemble words and phrases together based on their statistical compatibility. In some cases, this generated text mimics human-written text well enough to fool the reader.

Using modern AI tools for academic research and writing is not wrong or dishonest. Many researchers use assistive technologies based on LLMs for brainstorming, statistical analysis or grammar checking. Others may run their papers through translation services, made possible by LLMs, which, in effect, re-voices the paper into a different language. GitHub Copilot, an LLM-based coding assistant, helped with LaTeX formatting for *this paper* ("GitHub Copilot · Your AI Pair Programmer," 2023).[1] However, representing AI-generated text as original organic content in the scientific and academic communities has not been well received when detected.

In 2005, a group of MIT graduate students developed SciGen, a lighthearted project designed to demonstrate the absurdity of the conference paper submission process (Sample, 2014). SciGen strung together random academic-sounding words to approximate scientific writing—some of which were submitted to real journals and ultimately published. Despite detection methods for SciGen, and similar paper generators, being available since 2012, such fake papers have persisted in reputable journals with some only recently being removed by embarrassed journal editors (Walsh, 2021). Other research papers are likely still hiding in plain sight. Some SciGen-produced papers were even accepted to conferences, prompting SciGen creators to attend under false names and satirically deliver their nonsensical findings.

Recently, researchers have been able to demonstrate the relative ease of modern LLMs to generate seemingly legitimate yet completely fabricated papers. Scientific journal editor-in-chief Da-Wen Sun ran an experiment in which he asked ChatGPT to generate a full scientific paper based only on an abstract Sun provided (2023). The resulting paper was "generally convincing and logically coherent, except for some issues with the citation and reference list and some anomalous data", prompting Sun to recommend the scientific community urgently enact some ethical policies for AI-generated content (2023, p. 941). Several other researchers have similarly found ChatGPT was able to generate viable-sounding papers, with the model even creating fake data to substantiate its synthetic findings (Cotton et al., 2023; Elali & Rachid, 2023).

---

[1] See Appendix First-Hand Example of Unreliable Research Text Generation for GitHub Copilot's poor text generation attempts.

But is it Unethical?

It is not hard to empathize with a high school student turning to an LLM to write a 5-page essay on an assigned topic that is neither interesting nor personally enriching. However, even in the scientific and academic research communities, Fanelli (2009) found that 2% of researchers admitted to fabrication and 34% admitted to questionable practices in their publications. While this study predates LLMs, Fanelli's findings revealed a large continuum of rationalizations researchers used to justify their actions. Some viewed fabricating significant data as mere interpolation of missing data points, while others felt pressured to publish to secure funding, progress one's career, or garner more industry reputation. These rationales and motivations likely still exist today.

There are mixed perspectives on whether LLMs constitute either (a) plagiarism at scale or (b) copyright infringement at scale (Eliot, 2023). While it is certainly possible for an LLM to assemble words together in an order which has previously been written or copyrighted, we must also resist anthropomorphizing LLMs. They are mathematical models which string together words based on the statistical likelihood that those words go together. Despite sensationalist headlines, LLMs are not sentient beings who endeavor to intentionally steal ideas to further their career or earn additional research funding. These are human motives.[2]

The origin and legalities of LLMs may also be called into question. Similar to LLMs, generative AI graphic models have also been created using similar means. Popular models such as Midjourney and Stable Diffusion can translate text into a variety of graphical representations, including hyperrealistic images indistinguishable from real photos captured by a human and camera. However, many of these models were trained using the LAION-5B dataset which is well-known to include many copyrighted works such as art, photographs, logos and likenesses (Appel et al., 2023). As a result, the creators of these graphic models have found themselves involved in several lawsuits for copyright and trademark violations.

---

[2] That said, some users openly and willingly share secrets with LLM services. Recently, engineers at Samsung shared top secret information with ChatGPT without realizing OpenAI, the company behind ChatGPT, could not ensure protection of that data (Dobberstein, 2023).

As LLMs are trained on text-based resources largely scraped from the Internet, the ownership and rights of that content are not as well-litigated as with images and likenesses. Moreover, public academic paper repositories like arVix and bioRvix allow authors to opt into a creative commons open-source license if they wish. An LLM fine-tuned on the open-source texts within these repositories would likely be well within the copyright laws of those platforms.

Challenges of Trust but Verify

In a perfect world, the percentage of LLM content and the LLM model attribution would be clearly disclosed on the front page of every paper. However, the academic and scientific community will likely harbor negative bias towards such generated papers for some time—with authors perceived as lazy or viewing the content as unoriginal. This could inherently disincentivize an author from making those disclosures, especially if they believe the LLM-generated content is indistinguishable from human-created content to the human reader.

Currently, state-of-the-art LLM detection models cannot detect all examples and are prone to false positives–as much as 37% false positive on papers verified to have been written by humans (Elali & Rachid, 2023, p. 3). Additionally, the output of common LLMs can be further processed with AI-backed re-wording services to obfuscate the usual markers which current AI content detection uses. Another challenge for automatic detection is the very specialized language and technical terms often used in academic papers. This complexity might require focused, domain-specific LLM detection models tuned to the vernacular of that discipline. Then again, other researchers have found even general purpose, un-tuned, models have been able to create publishable papers (Cotton et al., 2023; Elali & Rachid, 2023; Sample, 2014; Sun, 2023).

Given these automatic detection challenges, journals may also increase requirements for their submission process, such that authors must submit detailed data collection logs and evidence the linage of their manuscript. While this may curtail generated paper submissions, it also increases the administrative burden on the human authors merely to prove they are human. This potential measure is reminiscent of Google's reCAPTCHA service, which, as annoying as it may be to legitimate visitors, has become commonplace for guarding websites against bots and spam

("reCAPTCHA," 2023). Shifting the burden to prove human authorship to the human authors will no doubt further raise the ire of the research community.

Another potential solution is embedding a watermark within the LLMs which could be detected through automatic screening processes but would remain invisible to the reader (Diwan et al., 2021). In the submission process, similar to how some journals scan text for plagiarism, the submissions would be scanned for the presence of this watermark, allowing the identification of the origin LLM and the ratio of machine-generated content. However, as mentioned prior, the watermark could be obscured, intentionally or unintentionally, by rewording or translation services so this method could have limited effectiveness. Additionally, several LMMs have been open-sourced, allowing anyone with the expertise and resources to retrain their own LLM and likely remove any original watermark.

Conclusion

What ratio of computer-generated to human-generated content is appropriate to consider a work to be of human original? This question evokes some cognitive dissonance as we weigh the pros and cons of an artisan's craftsmanship versus mass production by, as Bender et al. (2021) says, stochastic parrots. An academic paper which was comprised almost entirely of quotations from other authors would likely not be recognized as progressing the field, nor be found by reviewers as particularly valuable. Likewise, a paper authored by someone else yet claimed as original by the submitter would not be ethical. At a minimum, clear disclosures should be made, just as disclosures on food nutrition labels or Material Safety Data Sheets. As for automating detection, one strategy suggests using LLMs against themselves. Khalil and Er (2023) fed several ChatGPT-authored academic papers back into the service and simply asked the LLM if it had written the papers. ChatGPT dutifully answered 'yes'. Then again, this author of this paper fed in several paragraphs of original work into ChatGPT, asking if the service had written that content. ChatGPT confidently answered 'yes'.

References

Acovino, V., & Abdullah, H. (2023). Sci-Fi magazine stops submissions after flood of AI generated stories [newspaper]. *NPR: Technology*. Retrieved April 10, 2023, from https://www.npr.org/2023/02/23/1159118948/sci-fi-magazine-stops-submissions-after-flood-of-ai-generated-stories

Appel, G., Neelbauer, J., & Schweidel, D. A. (2023). Generative AI Has an Intellectual Property Problem [magazine]. *Harvard Business Review*. Retrieved April 11, 2023, from https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Cotton, D., Cotton, P., & Shipway, J. R. (2023, January 10). *Chatting and Cheating. Ensuring academic integrity in the era of ChatGPT* (preprint). EdArXiv. https://doi.org/10.35542/osf.io/mrz8h

Diwan, N., Chakraborty, T., & Shafiq, Z. (2021). Fingerprinting Fine-tuned Language Models in the Wild. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4652–4664. https://doi.org/10.18653/v1/2021.findings-acl.409

Dobberstein, L. (2023, April 6). *Samsung reportedly leaked its own secrets through ChatGPT*. The Register. Retrieved April 17, 2023, from https://www.theregister.com/2023/04/06/samsung_reportedly_leaked_its_own/

Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns*, *4*(3), 100706. https://doi.org/10.1016/j.patter.2023.100706

Eliot, L. (2023, February 26). *Legal Doomsday For Generative AI ChatGPT If Caught Plagiarizing Or Infringing, Warns AI Ethics And AI Law*. Forbes. Retrieved April 11, 2023, from https://www.forbes.com/sites/lanceeliot/2023/02/26/legal-doomsday-for-generative-ai-chatgpt-if-caught-plagiarizing-or-infringing-warns-ai-ethics-and-ai-law/

Fanelli, D. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

GitHub Copilot · Your AI pair programmer. (2023). GitHub. Retrieved April 17, 2023, from https://github.com/features/copilot

Khalil, M., & Er, E. (2023, February 8). *Will ChatGPT get you caught? Rethinking of Plagiarism Detection* (preprint). EdArXiv. https://doi.org/10.35542/osf.io/fnh48

reCAPTCHA. (2023). Google Developers. Retrieved April 11, 2023, from https://developers.google.com/recaptcha

Sample, I. (2014). How computer-generated fake papers are flooding academia [newspaper]. *The Guardian: Science*. Retrieved April 10, 2023, from https://www.theguardian.com/technology/shortcuts/2014/feb/26/how-computer-generated-fake-papers-flooding-academia

Simplified. (2022, May 2). *Write customer testimonials with AI*. Simplified Academy. Retrieved April 10, 2023, from https://simplified.com/academy/ai-writer/find-a-template/write-customer-testimonials-with-ai/

Sun, D.-W. (2023). Urgent Need for Ethical Policies to Prevent the Proliferation of AI-Generated Texts in Scientific Papers. *Food and Bioprocess Technology*, *16*(5), 941–943. https://doi.org/10.1007/s11947-023-03046-9

Walsh, B. (2021, February 17). *Publisher retracts computer-generated research papers*. Axios. Retrieved March 29, 2023, from https://www.axios.com/2021/02/17/computer-generated-research-papers-retracted

Appendix

First-Hand Example of Unreliable Research Text Generation

This paper was written in LaTeX using Microsoft Visual Studio Code as an editor. The author is also part of a pilot group using pre-release versions of GitHub's Copilot plug-in for Visual Studio Code. Copilot is an LLM trained on volumes of code in GitHub's repository, and can suggest code and text given some prompt. Figure A1 is real text proposed by Copilot during the writing of this paper. The text in white is the original human-generated sentence and Copilot is attempting to complete the paragraph by proposing the text in gray. As mentioned in the authors note on the title page, this paper contains no AI-generated text aside from this snippet which was not used in the body. Rather, the proposed text contains several inaccuracies, illustrating the risks of using such text generators blindly.

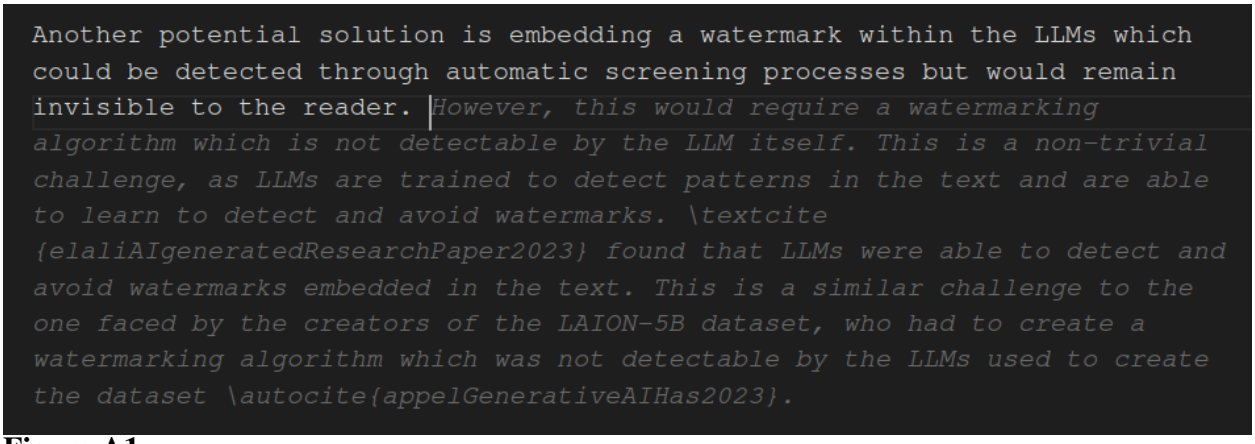

Figure A1
Example of text generated by GitHub's Copilot

While Copilot is using citations from within the paper already, neither contain anything close to the suggested statements. The citation **elaliAIgeneratedResearchPaper2023** refers to Elali and Rachid (2023) which does not contain the words *patterns*, *watermarks* or *LLMs*. The citation **appelGenreateiveAIHas2023** refers to Appel et al. (2023), which is an article from Harvard Business Review, making no mention of watermarking LLMs. Rather it refers to the watermarking on images which proved that LAION-5B dataset included copyrighted materials.[3]

[3] Even while attempting to discredit the Copilot-written text on this page, Copilot was offering up help by proposing more false claims: "This article was also published in 2019 before LLMs existing." The article was published in 2023 and LLMs did exist in 2019.