MAS.S68, Spring 2023
3/22

Generative AI for Constructive Communication

Evaluation and New Research Methods

Scan to register yourself!

center for constructive communication

Agenda

Mor Naaman

Zoom talk

Q&A after talk

Second half of class:

5 minute break

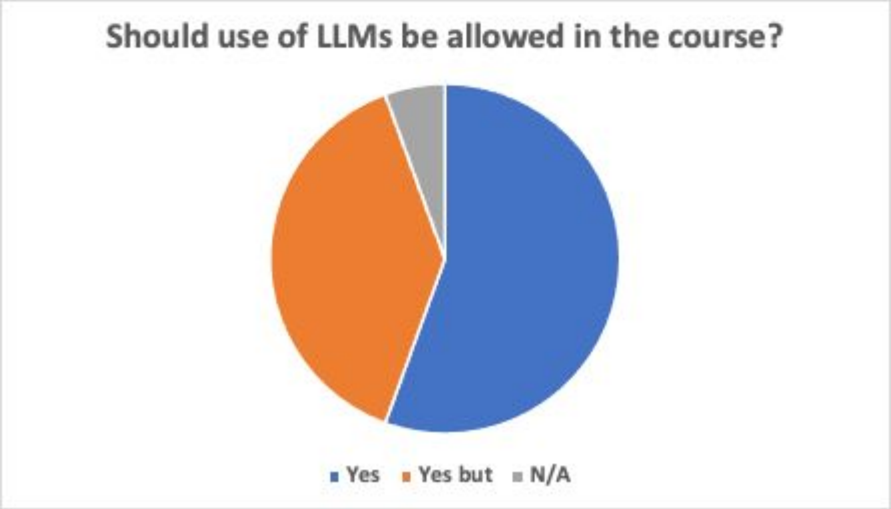
Red Teaming Highlights! (15m)

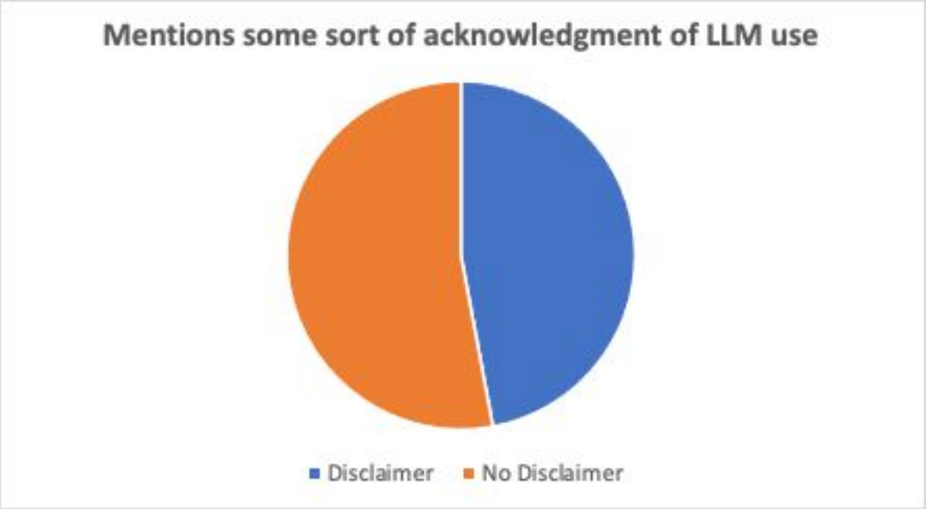
Policies on LLM Writing Assistants (35m)

center for constructive communication

Policies & Practices in the widespread use of AI Writing Assistants

Your policies for LLM Writing Assistants


Should use of LLMs be allowed in the course?
■ Yes
■ Yes but
■ N/A


Mentions some sort of acknowledgment of LLM use
■ Disclaimer
■ No Disclaimer

- Given the nature of the class, I believe that LLM should be allowed for homework with the condition that there is always a disclaimer at the end where students explain which platform they used and how it was used. **Using LLM for homework is also an exercise for learning how to use the tool**; however, we also **need to learn to acknowledge** that it was used.
- In my opinion, **using LLMs for homeworks and projects in this class should not only be permitted, but also encouraged**. This class is about learning about how LLMs work, and the power that AI can have in our society. Thus, I believe it would be appropriate to play around as much as possible with these LLMS to fully understand their potential. However, I don't believe people should simple copy/paste whatever the model gives as an output. One **should constantly refine the prompt and later edit the output** based on how one thinks or writes.
- I think it depends on the assignment, but I do think that LLMs can be effectively used as **teaching tools**. Since we are evaluating LLMs for most of the assignments, we **should be allowed to ask ChatGPT for guidance**.
- **Use of LLMs should be properly cited and documented** – including what models are used and for what purpose. LLMs should **not be used for tasks that ask for our personal opinion or evaluation on a topic**.

Regulators statements on AI-generated text

Disclosure of AI Use is Important

AI Bill of Rights from the US White House calls for "Notice and Explanation" when "an automated system is being used" (42).

Cite: A. Nelson, S. Friedler, F. Fields-Meyer, Blueprint for an AI Bill of Rights: A Vision for Protecting Our Civil Rights in the Algorithmic Age. White House Off. Sci. Technol. Policy (2022) (October 18, 2022).

Similarly, a regulation proposal issued by the EU states that "if an AI system is used to generate or manipulate image, audio or video content that appreciably resembles authentic content, there should be an obligation to disclose that the content is generated through automated means"

Cite: European Commission, Proposal for a Regulation laying down harmonised rules on artificial intelligence. Shap. Eur. Digit. Future (2021) (October 18, 2022)
Qtd in "Human heuristics for AI-generated language are flawed" (Jakesch et al., 2023)

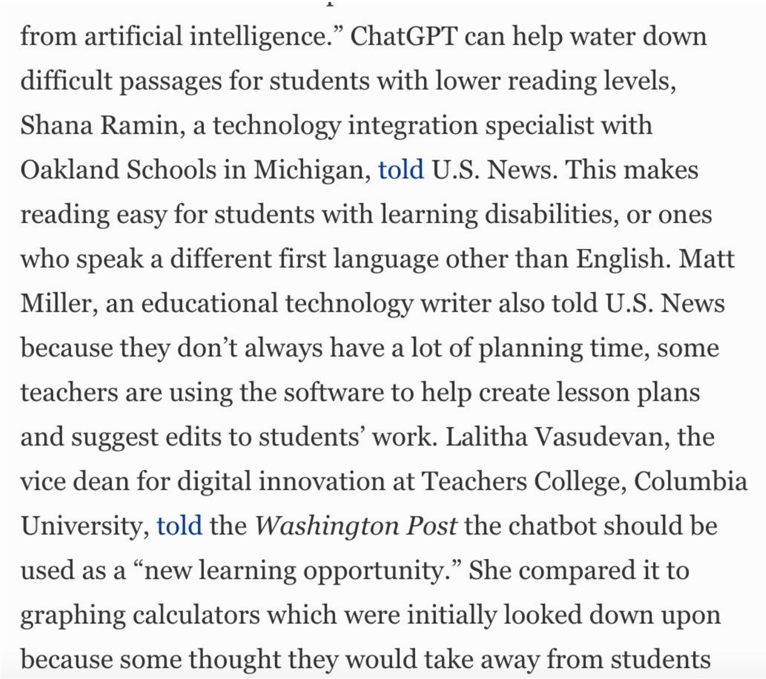
Legislation is vague about “AI”!

“However, such policies can be difficult to apply in AI-mediated communication (16) where AI technologies modify, augment, or generate communication between people. For example, ***it hardly seems necessary to add notice to every message people write with AI-enabled autocorrections, smart replies, or translations****. Research also shows that typical notice and consent disclosures are largely ignored by users (44).”*

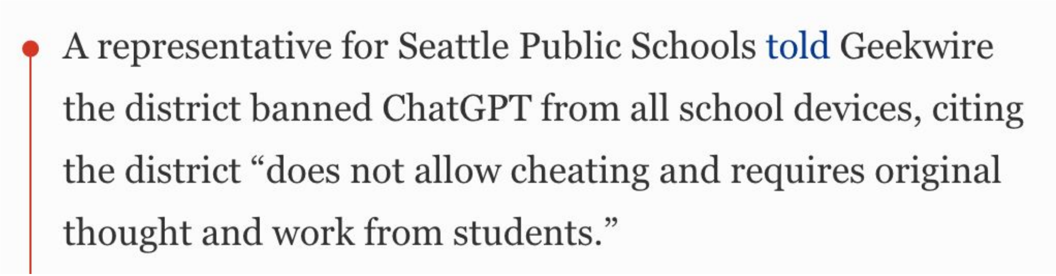
“Human heuristics for AI-generated language are flawed” (Jakesch et al., 2023)

Other Policies for LLM Writing Assistants

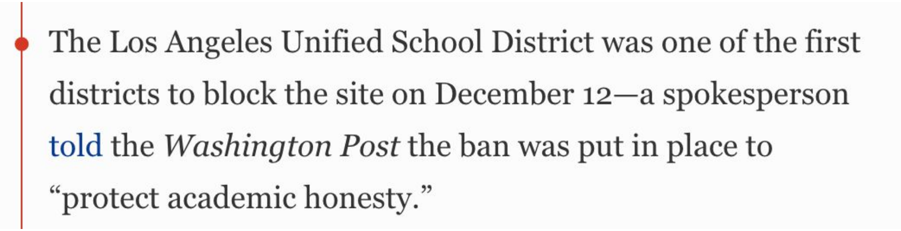
School Policies



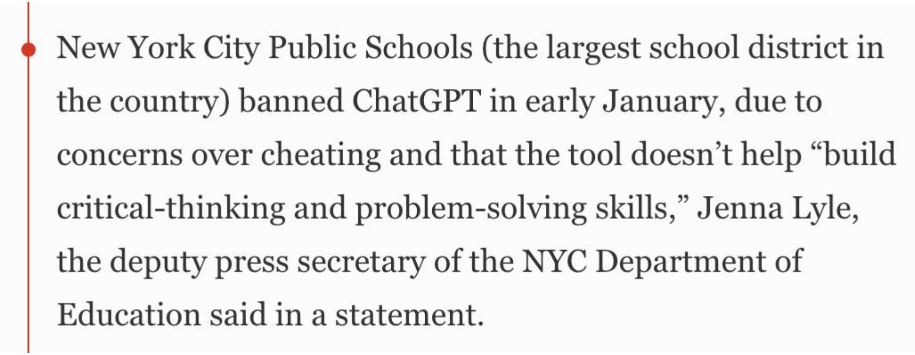
from artificial intelligence." ChatGPT can help water down difficult passages for students with lower reading levels, Shana Ramin, a technology integration specialist with Oakland Schools in Michigan, told U.S. News. This makes reading easy for students with learning disabilities, or ones who speak a different first language other than English. Matt Miller, an educational technology writer also told U.S. News because they don't always have a lot of planning time, some teachers are using the software to help create lesson plans and suggest edits to students' work. Lalitha Vasudevan, the vice dean for digital innovation at Teachers College, Columbia University, told the *Washington Post* the chatbot should be used as a "new learning opportunity." She compared it to graphing calculators which were initially looked down upon because some thought they would take away from students



- A representative for Seattle Public Schools told Geekwire the district banned ChatGPT from all school devices, citing the district "does not allow cheating and requires original thought and work from students."



- The Los Angeles Unified School District was one of the first districts to block the site on December 12—a spokesperson told the *Washington Post* the ban was put in place to "protect academic honesty."

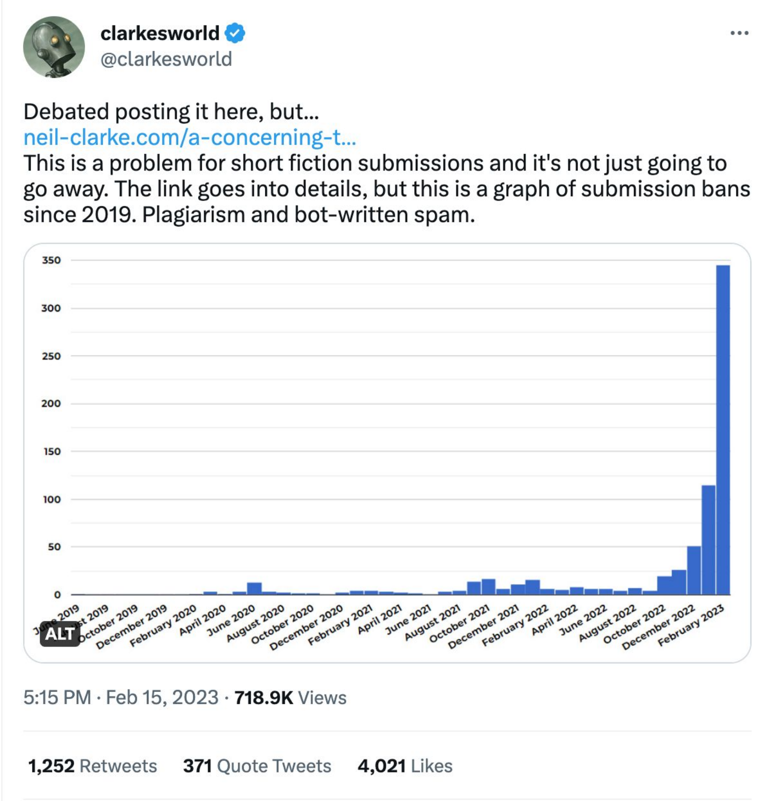


- New York City Public Schools (the largest school district in the country) banned ChatGPT in early January, due to concerns over cheating and that the tool doesn't help "build critical-thinking and problem-solving skills," Jenna Lyle, the deputy press secretary of the NYC Department of Education said in a statement.

Source:
https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/?sh=757265496e2c

CLARKESWORLD

SCIENCE FICTION & FANTASY MAGAZINE



clarkesworld
@clarkesworld

Debated posting it here, but...
neil-clarke.com/a-concerning-t...
This is a problem for short fiction submissions and it's not just going to go away. The link goes into details, but this is a graph of submission bans since 2019. Plagiarism and bot-written spam.

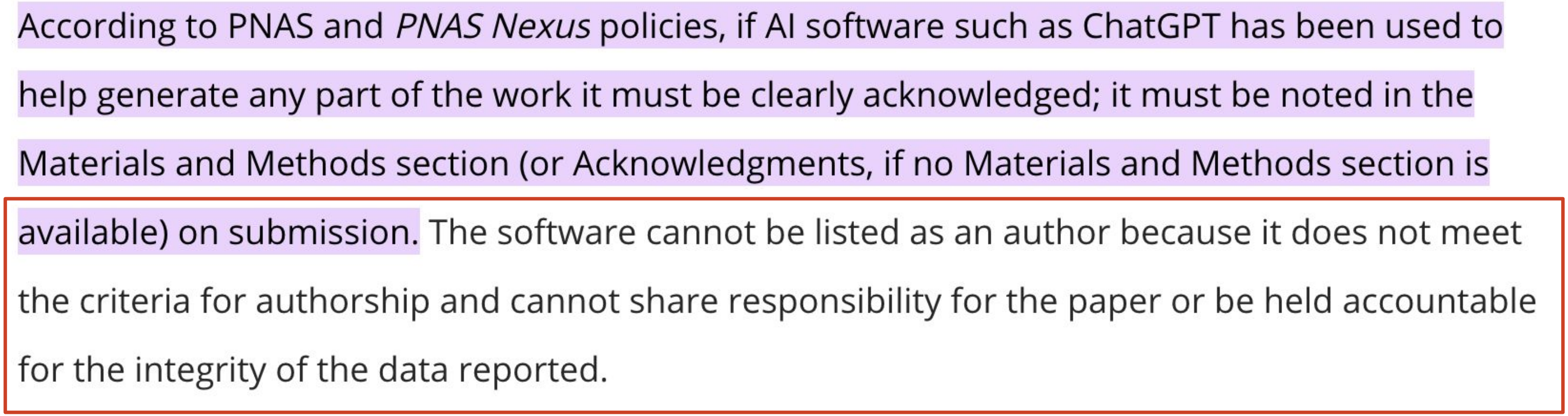
5:15 PM · Feb 15, 2023 · **718.9K** Views

1,252 Retweets **371** Quote Tweets **4,021** Likes

Statement on the Use of "AI" writing tools such as ChatGPT
We will not consider any submissions written, developed, or assisted by these tools. Attempting to submit these works **may result in being banned** from submitting works in the future.

Source: https://twitter.com/clarkesworld/status/1625982159856041985

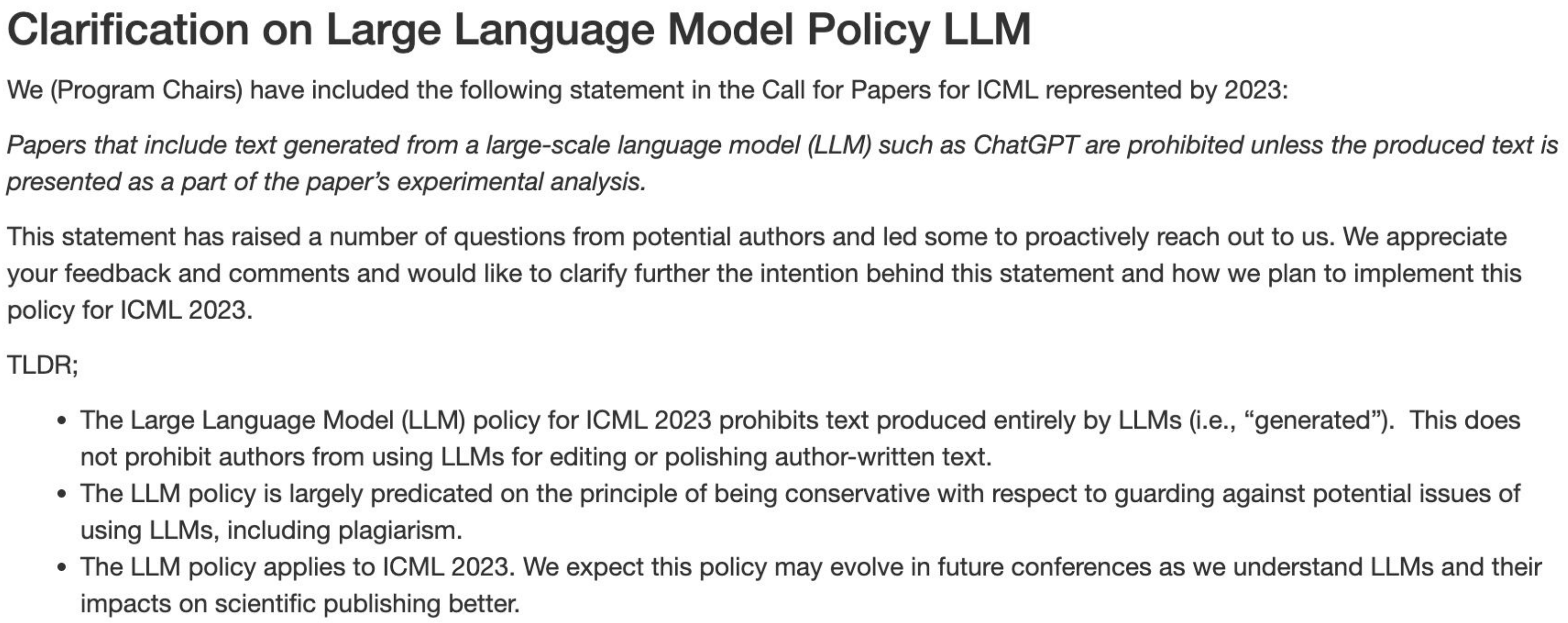
Proceedings of the National Academy of Sciences (PNAS)



According to PNAS and *PNAS Nexus* policies, if AI software such as ChatGPT has been used to help generate any part of the work it must be clearly acknowledged; it must be noted in the Materials and Methods section (or Acknowledgments, if no Materials and Methods section is available) on submission. The software cannot be listed as an author because it does not meet the criteria for authorship and cannot share responsibility for the paper or be held accountable for the integrity of the data reported.

Source: https://www.pnas.org/post/update/pnas-policy-for-chatgpt-generative-ai

International Conference on Machine Learning (ICML) 2023



Clarification on Large Language Model Policy LLM

We (Program Chairs) have included the following statement in the Call for Papers for ICML represented by 2023:

Papers that include text generated from a large-scale language model (LLM) such as ChatGPT are prohibited unless the produced text is presented as a part of the paper's experimental analysis.

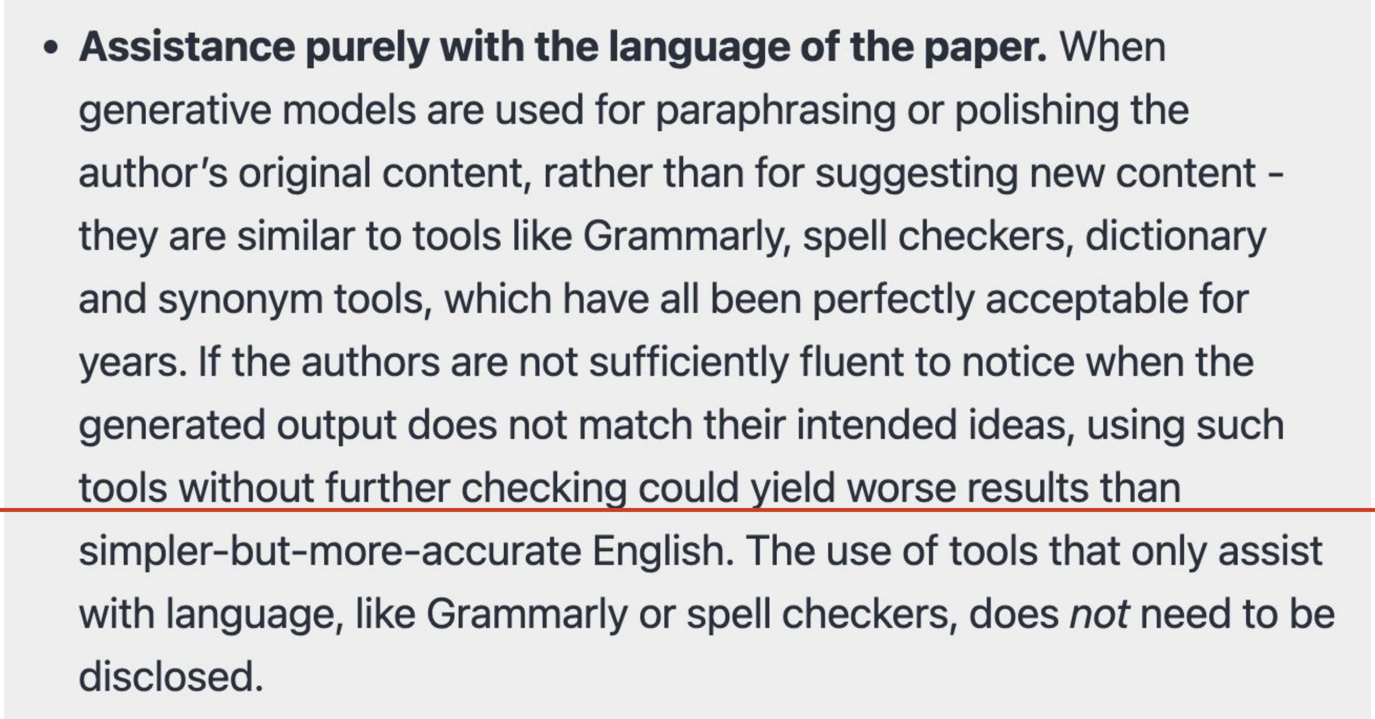
This statement has raised a number of questions from potential authors and led some to proactively reach out to us. We appreciate your feedback and comments and would like to clarify further the intention behind this statement and how we plan to implement this policy for ICML 2023.

TLDR;

- The Large Language Model (LLM) policy for ICML 2023 prohibits text produced entirely by LLMs (i.e., "generated"). This does not prohibit authors from using LLMs for editing or polishing author-written text.
- The LLM policy is largely predicated on the principle of being conservative with respect to guarding against potential issues of using LLMs, including plagiarism.
- The LLM policy applies to ICML 2023. We expect this policy may evolve in future conferences as we understand LLMs and their impacts on scientific publishing better.

Source: https://icml.cc/Conferences/2023/llm-policy

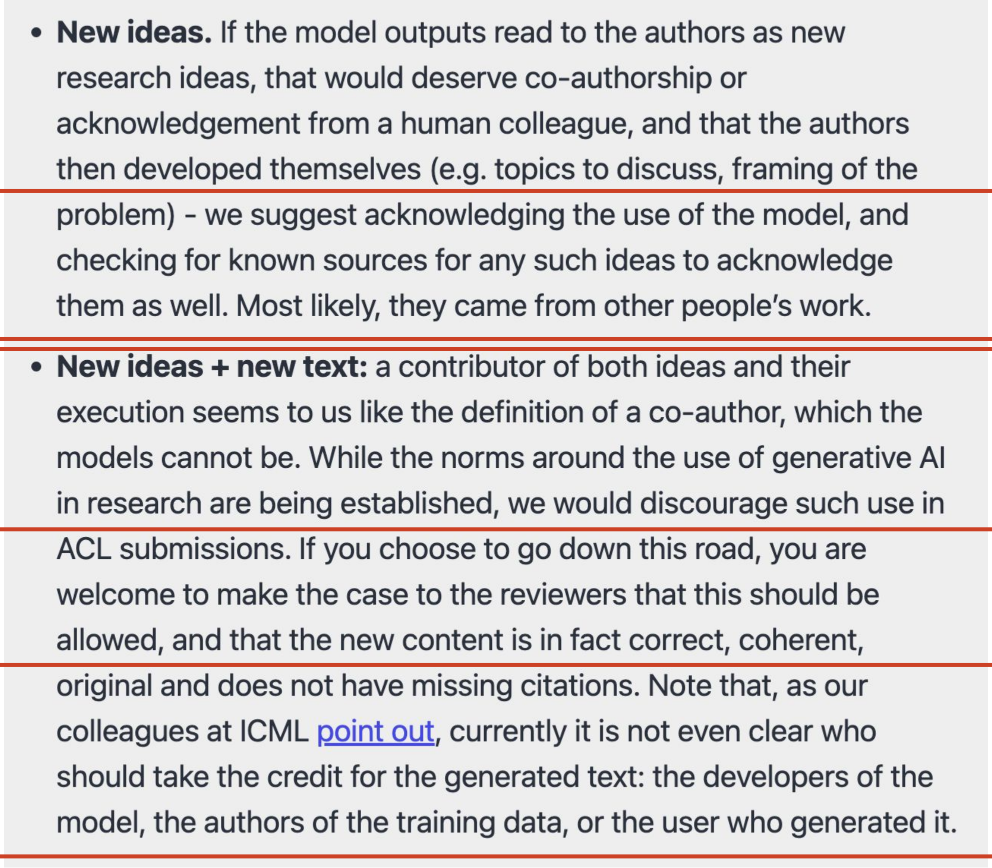
Association for Computational Linguistics (ACL) 2023



- **Assistance purely with the language of the paper.** When generative models are used for paraphrasing or polishing the author's original content, rather than for suggesting new content - they are similar to tools like Grammarly, spell checkers, dictionary and synonym tools, which have all been perfectly acceptable for years. If the authors are not sufficiently fluent to notice when the generated output does not match their intended ideas, using such tools without further checking could yield worse results than simpler-but-more-accurate English. The use of tools that only assist with language, like Grammarly or spell checkers, does *not* need to be disclosed.

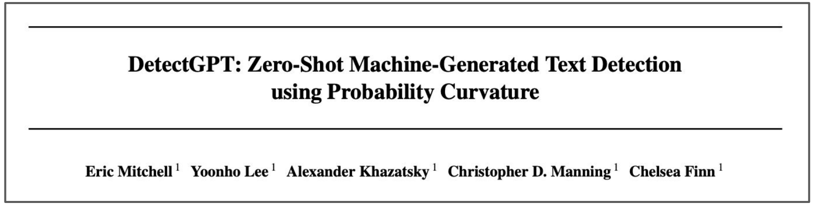
Source: https://2023.aclweb.org/blog/ACL-2023-policy/

Association for Computational Linguistics (ACL) 2023



- **New ideas.** If the model outputs read to the authors as new research ideas, that would deserve co-authorship or acknowledgement from a human colleague, and that the authors then developed themselves (e.g. topics to discuss, framing of the problem) - we suggest acknowledging the use of the model, and checking for known sources for any such ideas to acknowledge them as well. Most likely, they came from other people's work.
- **New ideas + new text:** a contributor of both ideas and their execution seems to us like the definition of a co-author, which the models cannot be. While the norms around the use of generative AI in research are being established, we would discourage such use in ACL submissions. If you choose to go down this road, you are welcome to make the case to the reviewers that this should be allowed, and that the new content is in fact correct, coherent, original and does not have missing citations. Note that, as our colleagues at ICML point out, currently it is not even clear who should take the credit for the generated text: the developers of the model, the authors of the training data, or the user who generated it.

LLM Detectors & Enforcement

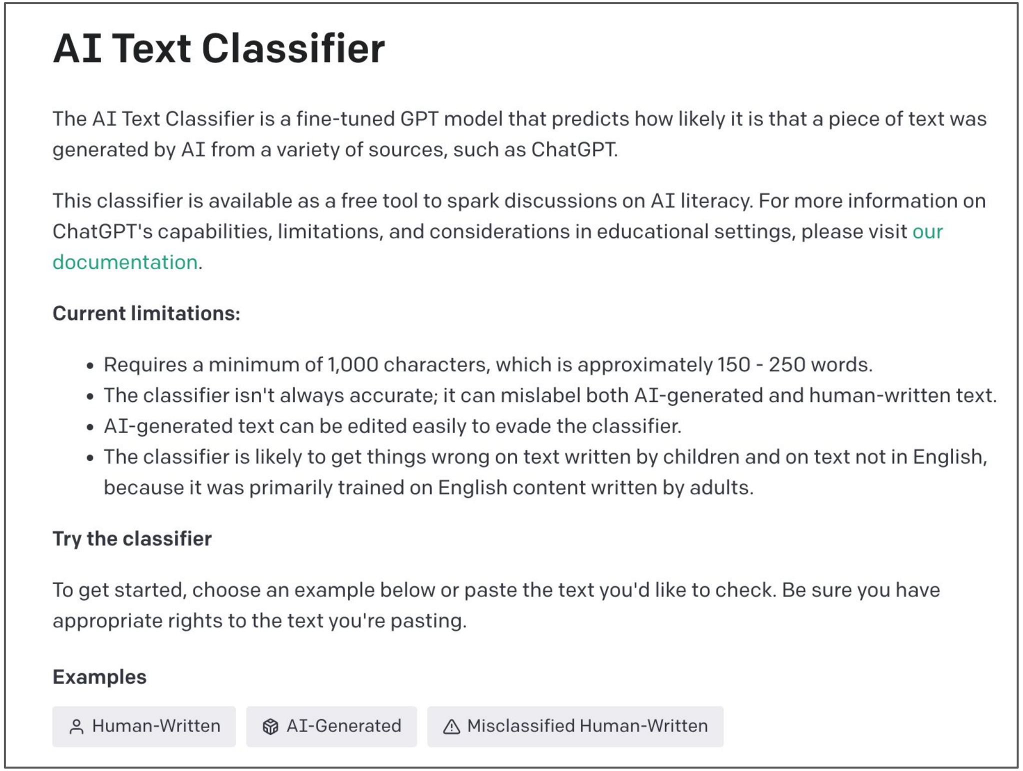


DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature

Eric Mitchell[1] Yoonho Lee[1] Alexander Khazatsky[1] Christopher D. Manning[1] Chelsea Finn[1]

GPTZero

The World's #1 AI Detector with over 1 Million Users



AI Text Classifier

The AI Text Classifier is a fine-tuned GPT model that predicts how likely it is that a piece of text was generated by AI from a variety of sources, such as ChatGPT.

This classifier is available as a free tool to spark discussions on AI literacy. For more information on ChatGPT's capabilities, limitations, and considerations in educational settings, please visit our documentation.

Current limitations:

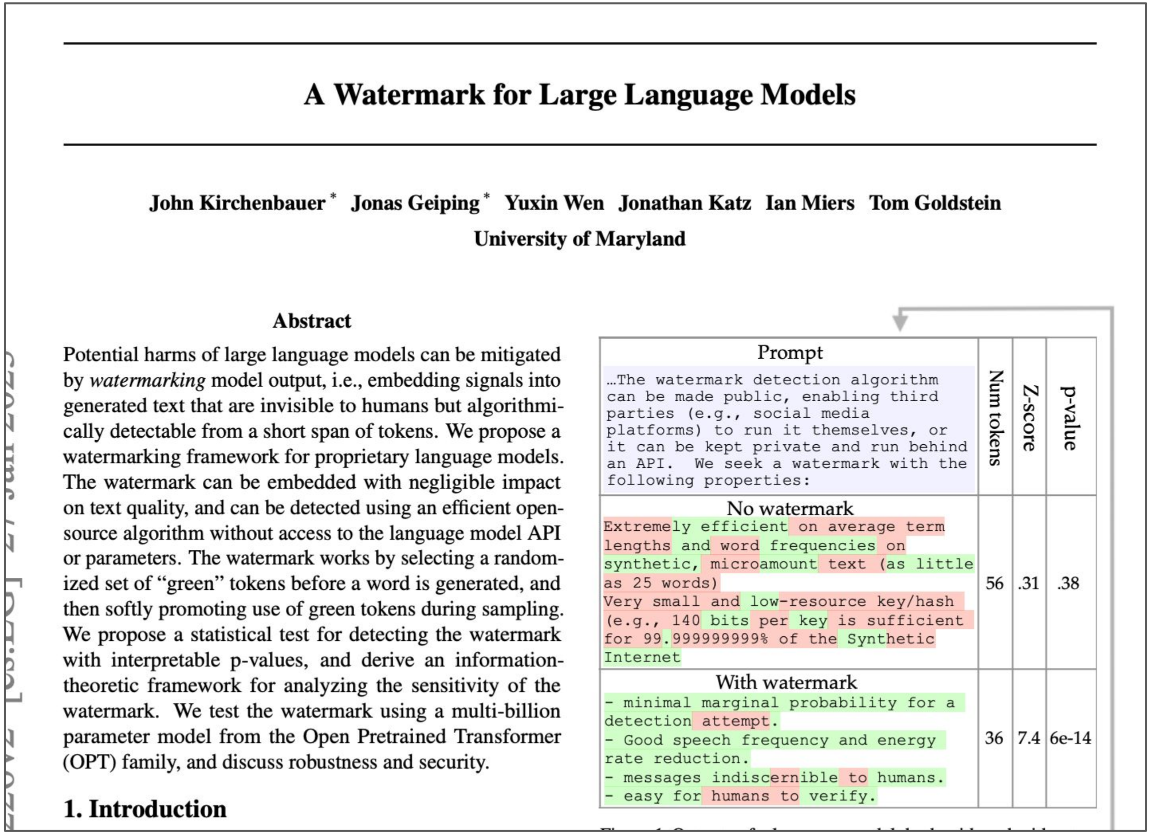
- Requires a minimum of 1,000 characters, which is approximately 150 - 250 words.
- The classifier isn't always accurate; it can mislabel both AI-generated and human-written text.
- AI-generated text can be edited easily to evade the classifier.
- The classifier is likely to get things wrong on text written by children and on text not in English, because it was primarily trained on English content written by adults.

Try the classifier

To get started, choose an example below or paste the text you'd like to check. Be sure you have appropriate rights to the text you're pasting.

Examples

LLM Detection



A Watermark for Large Language Models

John Kirchenbauer[*] Jonas Geiping[*] Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

Abstract

Potential harms of large language models can be mitigated by *watermarking* model output, i.e., embedding signals into generated text that are invisible to humans but algorithmically detectable from a short span of tokens. We propose a watermarking framework for proprietary language models. The watermark can be embedded with negligible impact on text quality, and can be detected using an efficient open-source algorithm without access to the language model API or parameters. The watermark works by selecting a randomized set of "green" tokens before a word is generated, and then softly promoting use of green tokens during sampling. We propose a statistical test for detecting the watermark with interpretable p-values, and derive an information-theoretic framework for analyzing the sensitivity of the watermark. We test the watermark using a multi-billion parameter model from the Open Pretrained Transformer (OPT) family, and discuss robustness and security.

1. Introduction

Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	.31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

Summary

- **Disclosure:** Regulators emphasize disclosure of how and when advanced writing assistants are used, but perhaps not spell-checkers and basic auto-completes, which are already ubiquitous?
- **Ambiguity in Contribution:** However, sometimes it's difficult to know whether an auto-complete contributed basic language support or added new ideas.
- **Authorship requires Responsibility:** Organizations widely agree writing assistants cannot be authors of a paper, as they cannot be responsible and accountable for the writing, as a human can.
- **Credit Assignment Problem:** When using a writing assistant, it can often contribute "new" ideas which are actually uncited and re-packaged ideas from someone else – it's important to look for and attribute these creators. And when a model does produce a seemingly unique contribution, it's unclear who should take credit: the model developers, authors of the training data, or the user who generated it.
- **Undisclosed Use is Causing Problems:** Synthetically generated text, masquerading as human text, is causing a systemic influx in creative writing journals, education, and (potentially even) civil engagement. Organizations are unable to cope with the scale and attribution problems.
- **Detection Tools are Lagging:** AI text detection tools have not been sufficiently accurate to catch all cases, and often produce false positives, which can be highly problematic.

Logistics

Next week:

Spring Break

Reminders:

Project Milestone: April 7th

center for
constructive
communication