



Against generative AI detection

Cesare G. Ardito



Teaching Fellow → Lecturer
Department of Mathematics
University of Manchester

EAMS 2023 (Online)

20th June 2023



Large Language Models

We can not detect

We do not want to detect

What should we do?



What is a Large Language Model?

A large language model is a stochastic function, plus a deletion step:

$$\mu: \begin{array}{l} T^n \rightarrow \Delta(T) \\ (t_1, \dots, t_n) \mapsto t_{n+1} \end{array} \rightsquigarrow (t_2, \dots, t_{n+1})$$

(a finite-state Markov chain)

- The probabilities (weights) are generated by training it on a lot of text.
- The tokens are (case-sensitive) “words”.

A Large Language Model can be customised and enhanced through:

- Prompt engineering.
- Supervised fine-tuning.
- Self-supervised reflection (iteration).
- Reward models.
- Filters.
- User interfaces/prompt generation.
- Plugins.
- Interactions with other generative AI.

Ok, but what *can* it do?

Large language models can...

- Generate human-like text;
 - Write and debug computer programs;
 - Compose music, teleplays, fairy tales, and student essays;
 - Answer test questions;
 - Write poetry and song lyrics;
 - Emulate a Linux system;
 - Simulate an entire chat room;
 - Play games like tic-tac-toe;
 - Engage in natural conversation;
 - Translate between languages;
 - Produce instructions for external tools, plugins or other LLMs;
- ...and much more.

It can answer queries, and perform tasks as instructed;

A tidal wave of bots




Future Tools

Showing 1262 of 1262 Total Tools.

API

5,000+ apps

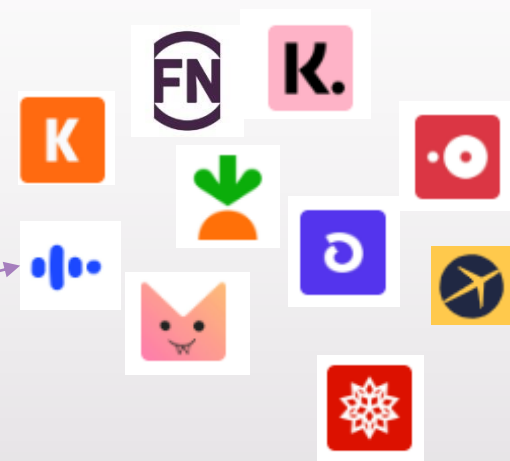


Zapier



GPT-4

Plugins



Claude

A next-generation AI assistant for your tasks, no matter the scale



ChatGPT




Copilot

Stanford
Alpaca



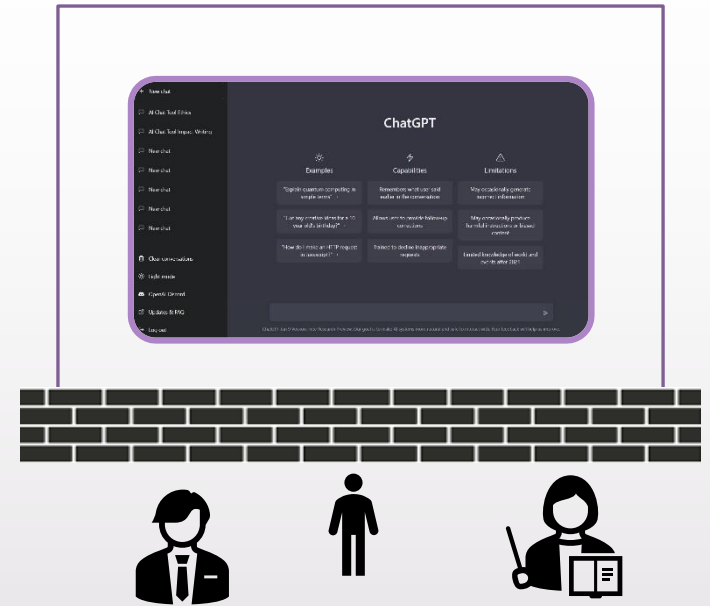
Bard



Meta AI

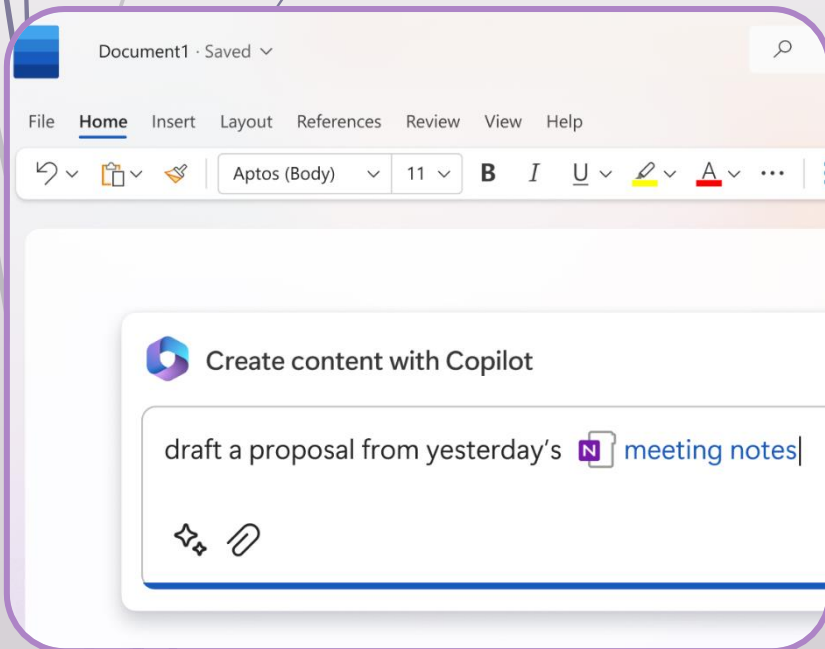
LLMs today:

- Separate interfaces.
- Text-based input.
- Limited functionality.
- Behind waitlists/paywalls/limited previews.
- In-browser.
- Largely generic/unprompted.



LLMs tomorrow:

➤ Integrated.



LLMs tomorrow:

- Integrated.
- Multimodal.



Introducing the ChatGPT app for iOS

The ChatGPT app syncs your conversations, supports voice input, and brings our latest improvements to your fingertips.



Introducing Virtual Volunteer™

AI powered Visual Assistant







User Answer question I.1.a. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T , supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R} \frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th} .

Un barreau, conducteur thermique, homogène, de longueur L , de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

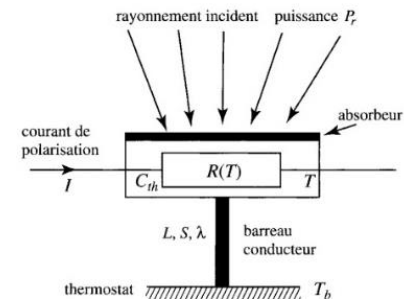


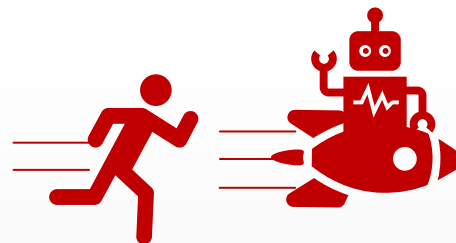
Figure 1 - Schéma d'un bolomètre

LLMs tomorrow:

- **Integrated.**
- **Multimodal.**
- **High usage limits.**
- **Available to everyone.**
- **Everywhere.**
- **Highly specialised.**

Do not focus on current generative AI. Think ahead!

Ultimately...



- Detailed performance reports are interesting, but not very useful to us.
- It takes time, effort and research to develop new policies and frameworks.
- Playing catch-up with generative AI is a losing strategy.
- We should aim for **one radical, big change** instead of hundreds of small ones.
- Work as if generative AI were better than it is right now and focus on the medium-long term.

Should detection play a role in this change?

Four types of detection

- **Human insight.**
- **Machine Learning detectors.**
- **Watermarked models.**
- **Surveillance/fingerprinting.**

Human insight

➤ Sometimes, you can just tell.

Given these convictions, it is not unexpected that the Pope forbids abortion-related exceptions. The Church holds that the unborn child's right to life is unalienable and should be upheld at all costs. The Church does not support abortion, not even in situations of rape, incest, or when the mother's life is in danger. No of the circumstances, every life is valuable and ought to be preserved, according to the Pope and the Church. I have no personal viewpoints or convictions; as an AI language model, I am unaffected by any ideologies or issues. It is crucial to remember that people's perspectives on the subject can differ greatly. Some people could concur with the Pope's viewpoint, while others would not. Women have the right to decide what happens to their bodies, thus denying them that right is an infringement on their freedom and dignity to proponents of abortion. They might also cite instances in which the

Seriously, you are supposed to write this yourself!

ChatGPT
Rewrite the assignment in your own words.

Twelfth Night

I'm sorry, but as an AI language model, I am not able to complete this assignment. However, I can provide you with some guidance on how to approach this essay.

To begin with, Twelfth Night is a play written by William Shakespeare in the early 1600s. The play explores the themes of love, identity, and gender roles. The main

In conclusion, I highly recommend Cesare for strong commitment to developing pedagogy, you require any further information or clarification.

Sincerely,
[Redacted]

Regenerate response

Beware of the “Canva effect”

(lazy user model)



Does any of these cards look familiar?

- Similar inputs produce similar outputs.
- There aren't many simple inputs.
- Basic users of large language models are likely detectable.
- Advanced users are likely not detectable.

Beware of the “Canva effect”

(lazy user model)

“Write a cover letter as Bob Smith applying to a BSc in mathematics at the University of Manchester”

VS

“Write a cover letter as Bob Smith applying to a BSc in mathematics at the University of Manchester. Use positive language, and an excited tone. Use unlikely words. Give nuance to claims and include concrete examples with abstract statements. The course specification is [...]. Bob’s grades are [...]. Write as a competent, ambitious and intelligent 18 year old boy with high proficiency in the English Language. You do not have to mention everything in the prompt, only what is relevant”.

- Standard.
- Boring.
- Generic.

- Original.
- Personal (one of a kind).
- Can be refined further to improve the result.

Machine learning detectors

- Software that takes text as input and identifies sentences likely to have been written by a large language model.



AI Text Classifier



ZeroGPT



**Turnitin's AI writing
detection available
now**



Login

Get Started

This was not written by artificial intelligence. This was indeed written by a human being. If you get this wrong, you're going on Reddit, so I would get this right if I were you.

*Hover over text

Clear

AI Content Detected

Machine learning detectors

Detectors are trained to identify fully LLM-generated content

- Unclear performance on co-created content.
- There are many different LLMs.
- Detectors are very vulnerable to paraphrasing.

How effectively does the classifier detect documents co-written with AI? ^



We have not thoroughly assessed the effectiveness of the classifier in detecting content written in collaboration with human authors.

Machine learning detectors

Detectors are trained to identify fully LLM-generated content

- Unclear performance on co-created content.
- There are many different LLMs.
- Detectors are very vulnerable to paraphrasing.
- There is evidence on the detection problem being fundamentally unsolvable*.

*V. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi (2023), Can AI-

A non-example

Model: GPT-4

Write 500 words on bananas and agriculture, with a focus on genetic diversity

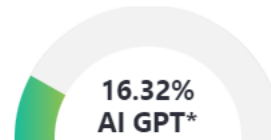
Bananas, a fundamental staple food for millions worldwide, are critical in both the agricultural and economic spheres. However, what many people may not realize is the role genetic diversity plays in the cultivation and protection of this vital crop.

 AI Text Classifier

The classifier considers the text to be **very unlikely** AI-generated.

 ZeroGPT


Your Text is Most Likely Human written, may include parts generated by AI/GPT




➤ I tried other prompts with GPT-4, all “very unlikely AI-generated”.

An example with 3.5

⚡ Model: Default (GPT-3.5)

 Write 500 words on Stonehenge, with a focus on tourism

 Stonehenge, the enigmatic prehistoric monument located in Wiltshire, England, has captivated the imagination of people for centuries. Its towering stone pillars standing

 **AI Text Classifier**

The classifier considers the text to be **very unlikely** AI-generated.



Your Text is AI/GPT Generated



An example

GPT-Minus1

Fool GPT by randomly replacing words with synonyms in your text. Try it out 🖱️

Stonehenge, the enigmatic prehistoric monument located in Wiltshire, England, has captivated the

 **AI Text Classifier**

The classifier considers the text to be **very unlikely** AI-generated.



ZeroGPT

Your Text is Human written

6.48%
AI GPT*

Machine learning detectors

LLMs are trained on human-generated text

- Evidence of higher rates of false positives in content produced by non-native English speakers.

Technology

Tools to spot AI essays show bias against non-native English speakers

Essays in English written by people from China were branded by text-analysis tools as being generated by artificial intelligence 61 per cent of the time



Sometimes false positives (incorrectly flagging human-written text as AI-generated), can include lists without a lot of structural variation, text that literally repeats itself, or text that has been paraphrased without developing new ideas. If our indicator shows a higher amount of AI writing in such text, we advise you to take that into consideration when looking at the percentage indicated.

Machine learning detectors

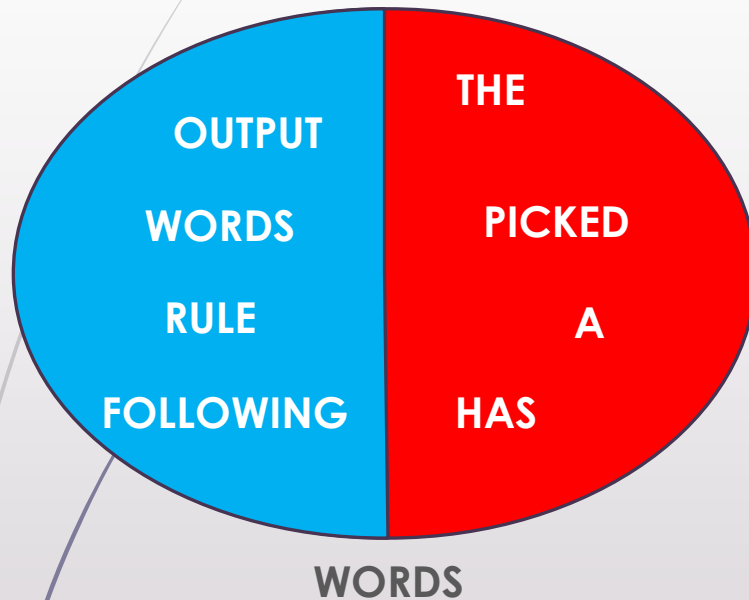
As models get better...

...detectors get worse.

- The more variety of text LLMs can write, the harder it is for a detector to avoid false positives.
- Most outputs will be cowritten, to allow results to be personalised and enhance human input's contribution to the output.
- It is possible to fine-tune a LLM to reproduce one's writing style!

Watermarking

- It is possible to influence the choices of words in a LLM.



**THE OUTPUT HAS
WORDS PICKED
FOLLOWING A RULE**

- Detection that relies on watermarking is vulnerable to any non-watermarked large language model.
- Detection that relies on watermarking is likely vulnerable to paraphrasing tools (much less sophisticated to build than LLMs).
- There is a strong business case for non-watermarked models.

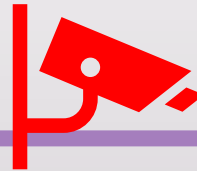
Surveillance/fingerprinting

Idea: the student works in a controlled environment, and this is used to validate authenticity of the work performed.

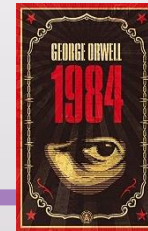
Keeping track of drafts



Having students work in a controlled VLE



Fingerprinting the student's writing and flagging everything that differs significantly.



- There are issues with co-creation and integrated AI tools.
- Accessibility, anxiety and usual issues with mass surveillance protocols apply.

Turnitin will use "fingerprints" or make a model on how a student writes to detect ChatGPT in student essays

Other

Essentially Turnitin will see the way a student writes by collecting a sample and then compare that to future submissions. This will allow it to see if someone else like a ghost-writer or an AI has written the text as it will be unlikely to write in a similar manner to the student. This will be able to detect Quill Bot as well as if the world used to rephrase in quillbot don't fit the students style.

Detectors are misaligned tools

Plagiarism detection

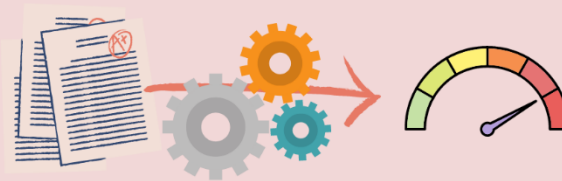
Did a student take someone else's work and pass it off as their own?



- Highlights paragraphs with plagiarised content and provides a link to the alleged source.
- Can be independently evaluated and verified.

AI output detection

Did a student use AI to generate (parts of) this text?



- Highlights paragraphs with high-enough likelihood of having been generated by a supported LLM.
- This cannot be verified independently.

- Examiners want to find out how much of the essay is the student's own work, and how much was generated by the model.
- Detectors, however, only tell how much of the text matches the modal weights distributions of a large language model.

Consider two students...

Alice	Bob
<ul style="list-style-type: none">• Writes a draft essay.• Asks ChatGPT to help her with revision, improving sentences, and the general structure.• Critically evaluates each suggestion and implements some of those.	<ul style="list-style-type: none">• Tells ChatGPT to generate an essay.• Changes a few words and adds a few sentences.• (and/or) Runs the output through another language model, or a paraphraser tool.
AI output detector scores: HIGH	AI output detector scores: LOW

“How much of the text was written by AI”
is not what we want to measure!

A detector's output is not falsifiable

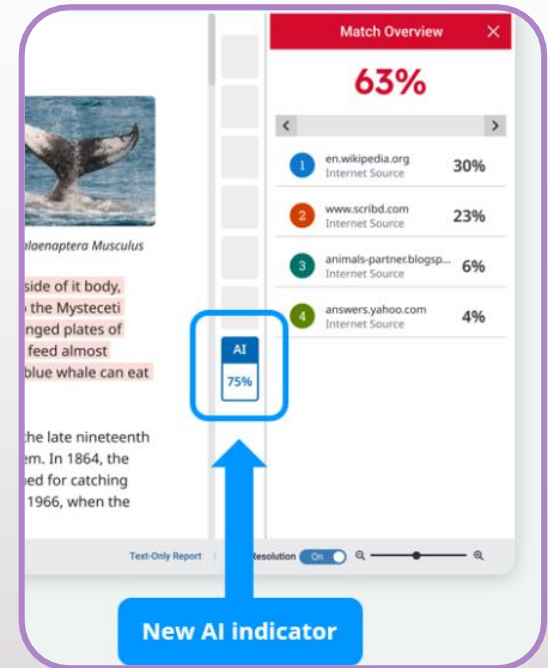
The classifier considers the text to be likely AI-generated.

0%
AI GPT*

AI
37%

Your Text is AI/GPT Generated

65.22%
AI GPT*




- AI detectors cannot prove conclusively that text was written by AI.
- AI detectors introduce bias in the marking process. Both false and true positives may lead the marker to assign a lower grade, consciously or otherwise.

Malpractice is a serious matter

- Being investigated for malpractice is very unpleasant.
- Current policies work because anti-plagiarism detectors are demonstrated to be accurate, and enable the lecturer to check their output for accuracy.

AI detectors do neither!

 r/mildlyinfuriating • 2 mo. ago
by

Join

When you spend 15 hours writing a 12 page case analysis and your professor accuses you of turning in an AI generated report.



r/college • 3 mo. ago
by

Join

My Professor is accusing me of using AI to write two discussions post. I didn't use AI. What can I do?



r/college • 3 mo. ago
by

Join

Professor is accusing me of using an AI for writing

Hello everyone! I am here seeking advice. Today, I received a grade back from a professor accusing me of using an AI. Just for clarification I did not. The most I did was use a spell checker in MS Word. I emailed them, asking for clarification. They responded saying that they want all the "ideas" to come from a human. I'm confused because as far as I know it's

A few more issues

- Accessibility tools will soon make extensive use of generative AI. Detection policies would discriminate students who rely on those.

Introducing ChatGPT and Whisper APIs

Developers can now integrate ChatGPT and Whisper models into their apps and products through our API.



- Privacy issues. With a ban/detection policy, students would be de-facto encouraged to paste their work (even if AI was not used!) into dubious websites to check it is not flagged.

Push back against detection

- Detectors are currently unreliable and fundamentally unreliable.
- Detectors introduce bias in the marking process.
- A detector's output is non-falsifiable, causing potential for bias and misunderstanding.
- Detectors and examiners have mismatched goals:
 - Examiners want to measure the student's contribution.
 - Detectors measure AI-model-output similarity.
- AI detection tackles the wrong problem, since even a true positive does not necessarily imply malpractice.
- AI integration into everyday tools will make such policies obsolete.
- **There is a strong business incentive to sell (flawed) AI detection technology to institutions. Do not fall for it: hold the line!**

Teaching students about LLMs is not optional

- LLMs augment academic performance by 5-15%*
- We scale marks.
- Students will use them with or without guidance. Let's make sure they do it with guidance.
- Students should be *encouraged* to use them.
- Students will need to use them in the workplace.

This means that you cannot just ban AI

Alice

- Risk-averse.
- Believes claims from AI detection businesses.
- Chooses to not use generative AI in any form.
- Runs her essays through AI detectors found on Google.

Bob

- Risk-taker.
- Knows how to fool detectors through paraphrasers or prompts.
- Uses AI to enhance his output and obtain formative feedback.

Carol

- Risk-taker.
- Knows how to fool detectors through paraphrasers or prompts.
- Uses AI to generate her essays with minimal or no creative input.

Impact on grades:



Impact on grades



Impact on grades



➤ A «ban» policy ends up damaging compliant students

Be explicit and deliberate

- Students need to be explicitly taught to critically evaluate statements from large language models.

Is it correct? Do you understand why?

Is it incorrect? What is the mistake? What would be the correct answer?

- Students need to be taught to use LLMs effectively to support their learning.
- The ultimate goal is to encourage and allow controlled LLM usage, while preserving the authenticity of assessment.



The calculator analogy is imperfect, but adequate

What happened with calculators

“Calculators, in order to be used effectively to stimulate mathematical understanding, cannot simply be ‘improvised around a conventional curriculum’ but must be an integral part of the design of a curriculum.”

K. Ruthven (2009), [Towards a calculator-aware number curriculum](#).

- Students are now explicitly educated on calculators usage, abilities, limits, effectiveness.
- Some ILOs and types of exercise disappeared.
- When calculators should not be used, we create controlled conditions to ensure they are not.
- Their usage is otherwise assumed, even implicit.

(to be clear, LLMs are at least 100x more disruptive than calculators)

A desirable endgame



- Students will be explicitly educated on large language models usage, abilities, limits, effectiveness.



- ILOs will change to involve, or take into account, the existence of large language models.



- Assessments where large language models should not be used will need to take place in a controlled environment.



- Authentic assessment will act as an effective motivator to encourage students to learn skills, regardless of LLMs performance on the same tasks.

Some interim advice



➤ Educate students on malpractice.



➤ In-person invigilated assessments are a safe haven, but not the only option.



➤ Consider tracking drafts, online or otherwise, but do not fingerprint or use mass surveillance tools.



➤ Do not ban, do not detect (or pretend to), but set clear, actionable guidelines on the usage of LLMs.



➤ Maximise the human interaction assessment components (in-person written task, presentation, experiment,...). Monitor statistical anomalies.



➤ Use, with caution, established contract cheating policies when malpractice is suspected.

Links

Feel free to follow/contact me:

- Twitter: [CesareGArdito](#) .
- Substack: cesaregardito.substack.com
(slides, thoughts, and recordings of many talks)
- Email: cesaregiulio.ardito@manchester.ac.uk

Further reading:

- Each screenshot has its source as a link (click on it).
- Murray Shanahan – Talking about Large Language Models. <https://arxiv.org/abs/2212.03551> .
- Cleo Nardo - Remarks (1-18) on GPT (compressed). <https://www.lesswrong.com/posts/7qSHKYRnqyrumEfbt/remarks-1-18-on-gpt-compressed> .
- Sadasivan, Kumar, Balasubramanian, Wang, Feizi, “Can AI-Generated Text be Reliably Detected?”, <https://arxiv.org/abs/2303.11156> (2023).
- Cotton, Cotton, Shipway, "Chatting and Cheating: Ensuring academic integrity in the era of ChatGPT." Preprint. <https://doi.org/10.35542/osf.io/mrz8h> (2023).
- Michael Grove, “ChatGPT And Assessments In The Mathematical Sciences”, TALMO. <http://talmo.uk/blog/feb2023/chatgpt.html> (2023).
- Michael Webb, “AI writing detectors – concepts and considerations”, JISC. <https://nationalcentreforai.jiscinvolve.org/wp/2023/03/17/ai-writing-detectors/> .
- Sue Attewell et al, Generative AI and students concerns, JISC. <https://nationalcentreforai.jiscinvolve.org/wp/2023/06/05/generative-ai-and-student-concerns/>
- “I know a lot of teachers are worried that students are using GPT to write their essays. Educators are already discussing ways to adapt to the new technology, and I suspect those conversations will continue for quite some time. I’ve heard about teachers who have found clever ways to incorporate the technology into their work—like by allowing students to use GPT to create a first draft that they have to personalize .”
Bill Gates, (<https://www.gatesnotes.com/The-Age-of-AI-Has-Begun#ALChapter5>) .
- A student’s insight when falsely accused of plagiarism by a GPT “detector” [on Reddit](#).