MANCHESTER
1824
The University of Manchester

Large Language Models and assessment

Cesare G. Ardito
Teaching Fellow
Department of Mathematics

These slides are meant to go together with the talk:
many things might not make sense from the slides only.
I suggest to watch the recording: [link]

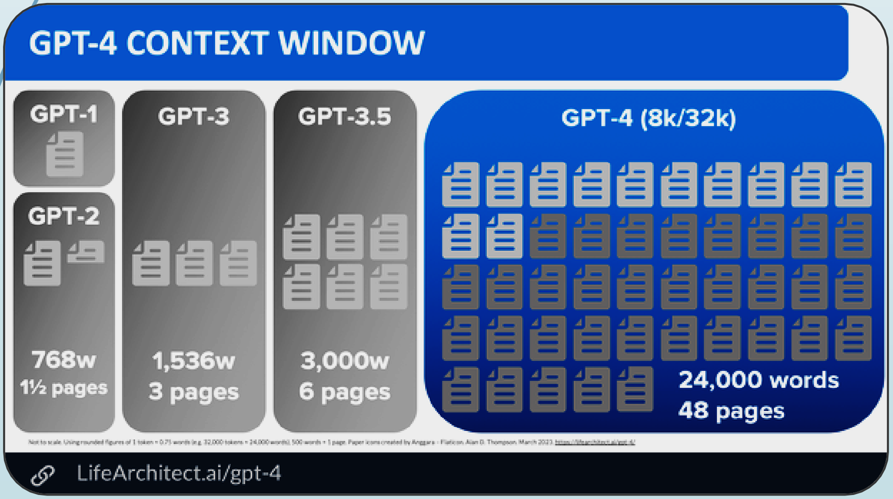
Large Language Models

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

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

(a finite-state Markov chain)

The probabilities (weights) are generated by training it on a big corpus of text.

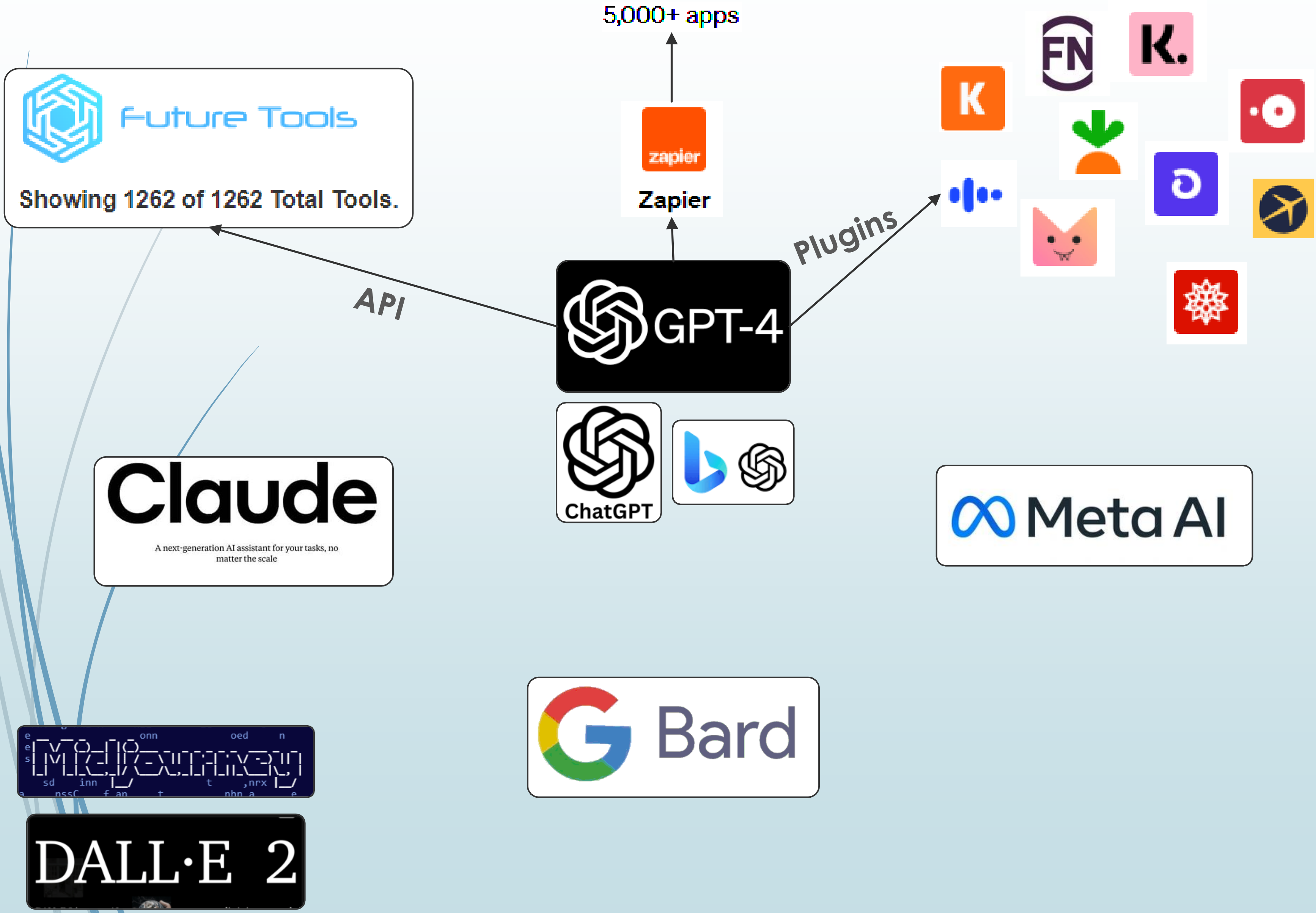


Almost exactly the memory capacity of the Commodore 64

A Large Language Model can be customised through:

- Prompt engineering.
- Supervised fine-tuning.
- Self-supervised reflection (iteration).
- Reward models.
- Filters.
- Access to tools.
- User interfaces/prompt generation.
- Plugins. NEW

A tidal wave of bots



5,000+ apps

Future Tools

Showing 1262 of 1262 Total Tools.

zapier

Zapier

API

Plugins

GPT-4

ChatGPT

Claude

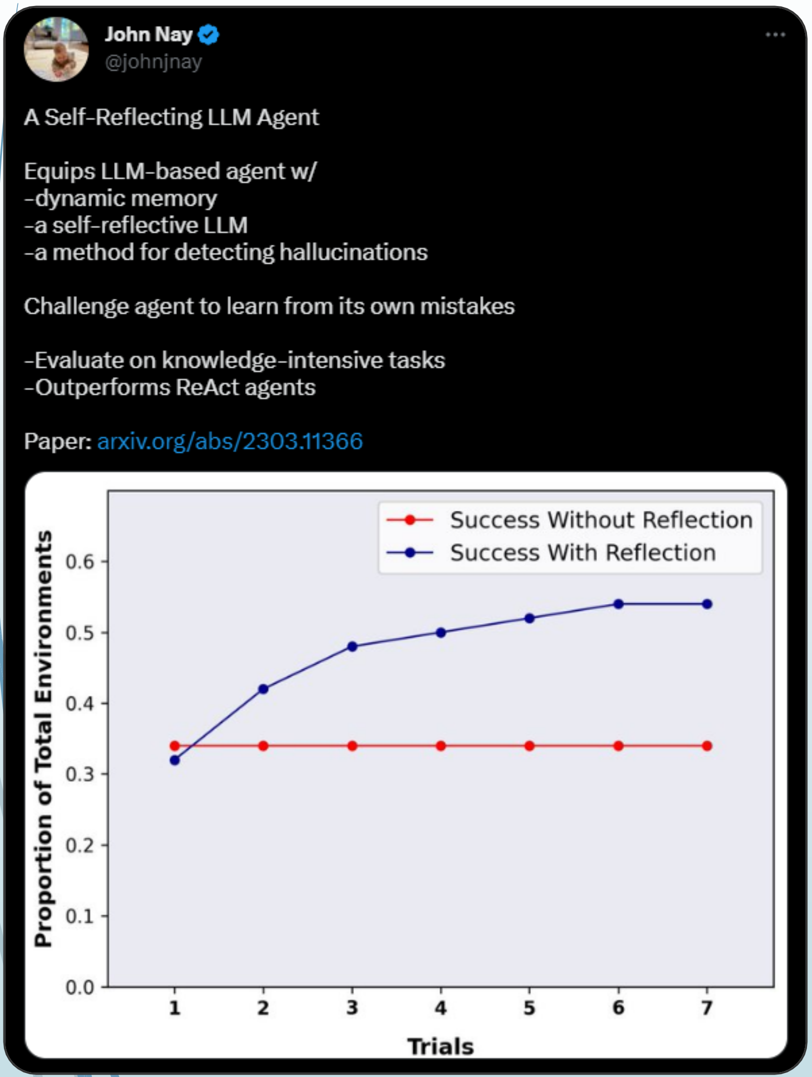
A next-generation AI assistant for your tasks, no matter the scale

Meta AI

Bard

DALL·E 2

Abilities of variations of the same model can be substantially different



John Nay
@johnjnay

A Self-Reflecting LLM Agent

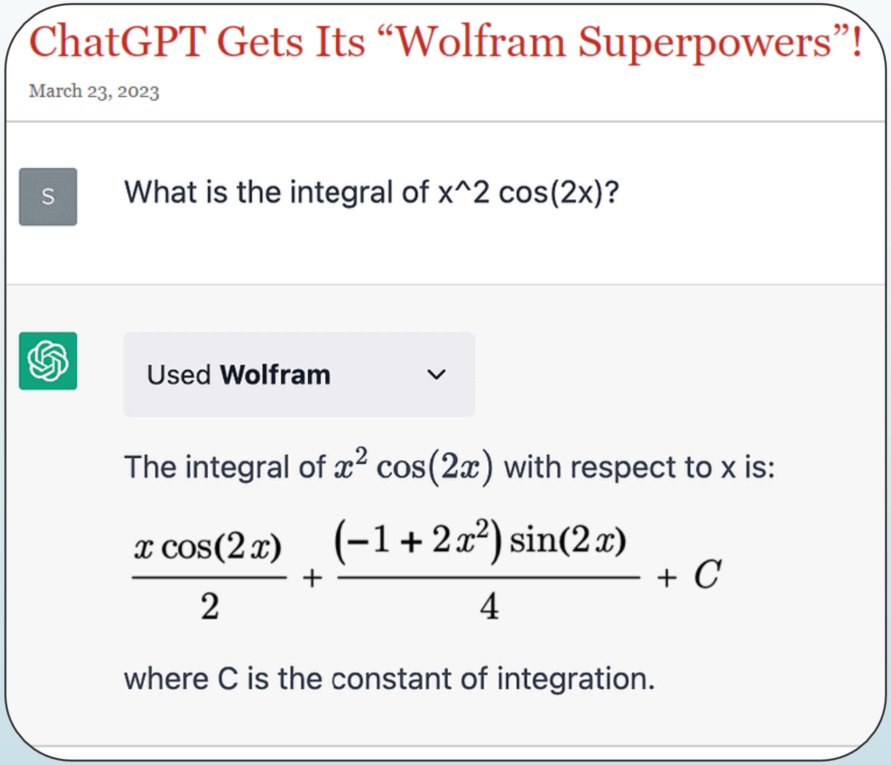
Equips LLM-based agent w/
-dynamic memory
-a self-reflective LLM
-a method for detecting hallucinations

Challenge agent to learn from its own mistakes

-Evaluate on knowledge-intensive tasks
-Outperforms ReAct agents

Paper: arxiv.org/abs/2303.11366

Self-supervised



ChatGPT Gets Its "Wolfram Superpowers"!

March 23, 2023

S What is the integral of x^2 cos(2x)?

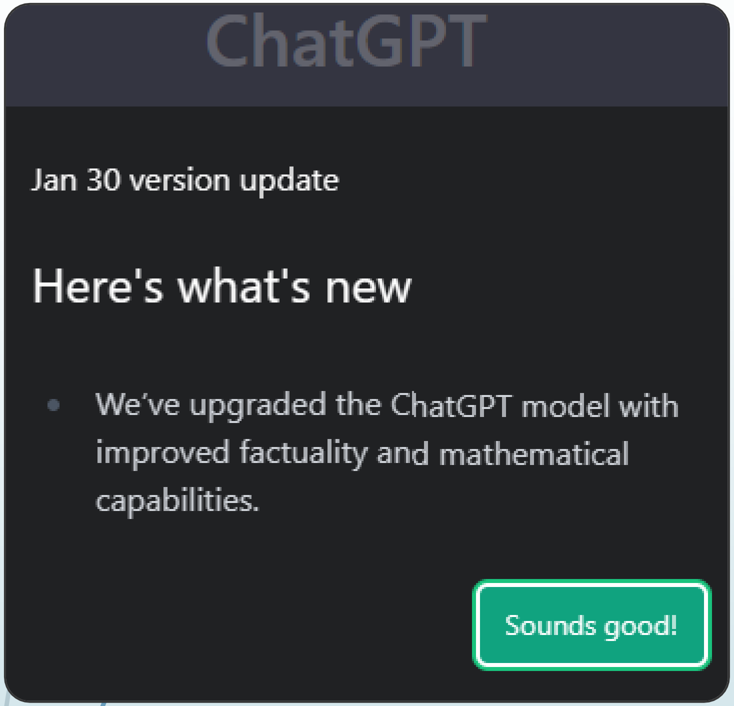
The integral of $x^2 \cos(2x)$ with respect to x is:

$$\frac{x\cos(2x)}{2} + \frac{(-1+2x^2)\sin(2x)}{4} + C$$

where C is the constant of integration.

Plugins

New, or updated models come out all the time



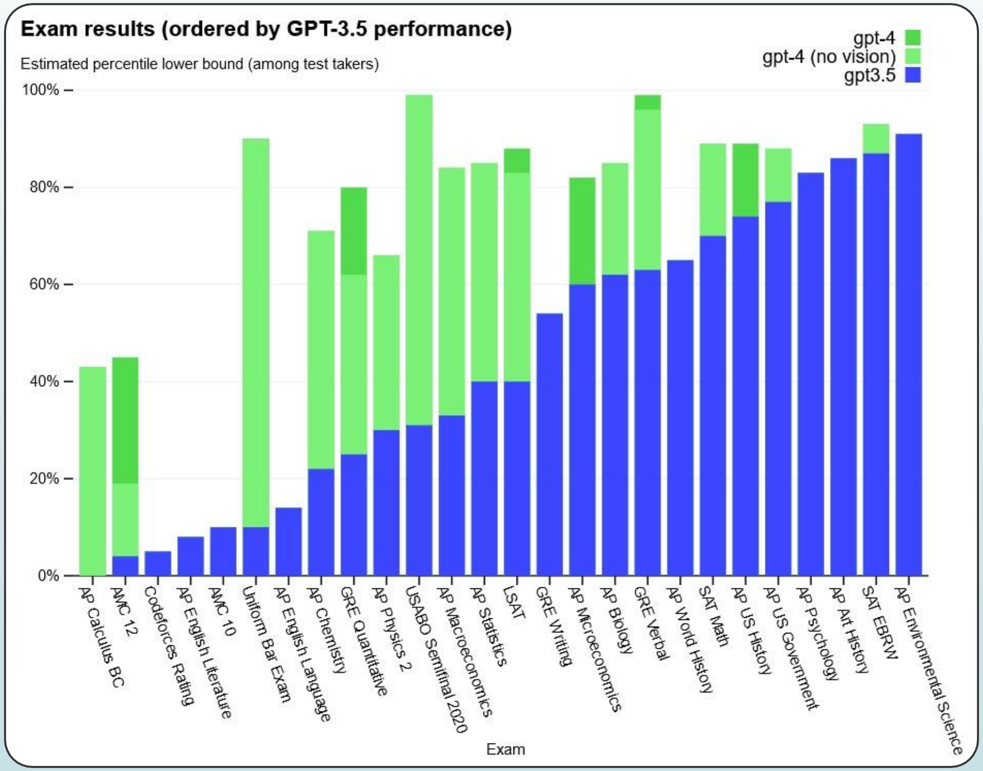
ChatGPT

Jan 30 version update

Here's what's new

- We've upgraded the ChatGPT model with improved factuality and mathematical capabilities.

Changelog of a GPT-3.5 update

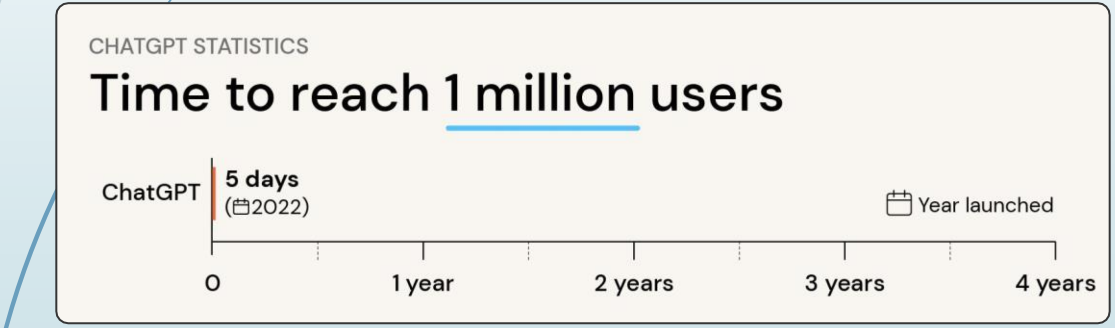


GPT-3.5 to GPT-4

Given a task X, can a Large Language Model do X?

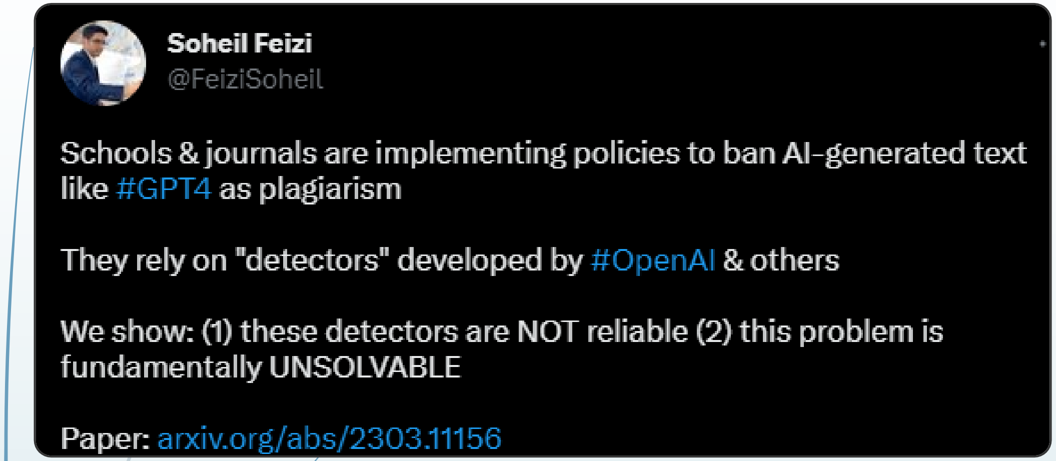
- For reasonable X, if the answer is not “Yes” then it is “Maybe”.

- A variation able to perform X could be discovered and implemented faster than we can possibly react.



- The pace of progress is such that even experts in the field struggle to keep up.

We can’t reliably detect LLM output



Soheil Feizi
@FeiziSoheil

Schools & journals are implementing policies to ban AI-generated text like #GPT4 as plagiarism

They rely on "detectors" developed by #OpenAI & others

We show: (1) these detectors are NOT reliable (2) this problem is fundamentally UNSOLVABLE

Paper: arxiv.org/abs/2303.11156

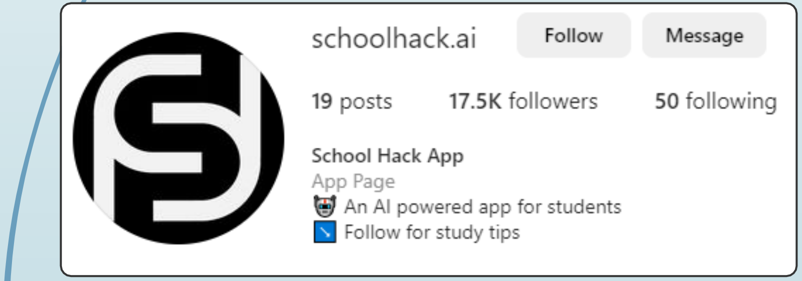
Can AI-Generated Text be Reliably Detected?

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi

https://arxiv.org/abs/2303.11156

Further:

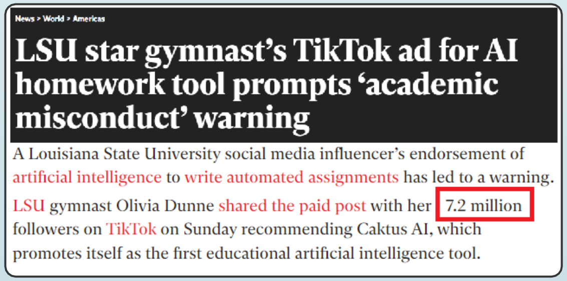
- **Even assuming a working detector can exist, as soon as a variation of an LLM is discovered to bypass it, it can reach every interested user in a very short time.**



schoolhack.ai

19 posts 17.5K followers 50 following

School Hack App
App Page
An AI powered app for students
Follow for study tips



News › World › Americas

LSU star gymnast’s TikTok ad for AI homework tool prompts ‘academic misconduct’ warning

A Louisiana State University social media influencer’s endorsement of artificial intelligence to write automated assignments has led to a warning.

LSU gymnast Olivia Dunne shared the paid post with her 7.2 million followers on TikTok on Sunday recommending Caktus AI, which promotes itself as the first educational artificial intelligence tool.

- **Traditional “laziness” of plagiarists cannot be relied upon. Every sophisticated approach can simply be automated.**

- **False positives rates are high, and accusing someone of plagiarism is serious.**

My proposal

Assumption: ***With minimal input and effort the LLM can perfectly, instantly and freely simulate a good student's output.***

- This is not yet true. But it is easier.

- However, successful teaching and assessment policies will be those that would still work in a world where this is true.

- It takes time, effort and research to develop good policies. Choose one radical, big change instead of hundreds of small ones.

- The alternative is continuously redesigning policies readapting everything to the ever-changing frontier of LLM abilities. New things come out every day.

A few possible strategies

Educate students on plagiarism.

Track drafts, or require students to work on online platforms that can track input and the progress of the work.

Do not ban, but set clear guidelines on the usage of LLMs.

Maximise the human-to-human interaction assessment components, ideally including at least one in every course (in-person written task, presentation, experiment,…). Monitor statistical anomalies.

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

Final thoughts

Several “AI detectors” companies will claim to be able to reliably detect AI-generated output. Do not fall for this trap. Hold the line.

Engage with LLMs: use ChatGPT! Learn prompts! Run your assessments in it! Follow people experimenting with it!

It will be extremely important to preserve human critical thinking. To do this, it is important we can assess students accurately.

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

Hybrid assessments that include in their design the possible usage of LLMs could help us prepare students for the new world they will face.

Opinions all mine ☺

Thank you for your attention

Feel free to contact me:

- Twitter: CesareGArdito .
- Substack: cesaregardito.substack.com

  (slides, thoughts, and recording of this and other talks)
- Email: cesaregiulio.ardito@manchester.ac.uk

Sources/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 .
- **Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, Soheil Feizi, "Can AI-Generated Text be Reliably Detected?",** https://arxiv.org/abs/2303.11156 **(2023).**
- **Cotton, Debby RE, Peter A. Cotton, and J. Reuben 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).**
- *"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.