A blue arrow pointing right is located at the top left. Below it, several thin, grey, curved lines extend downwards and to the right, creating a decorative background element.

Recent developments for large language models and mathematics education

Cesare G. Ardito

Lecturer

Department of Mathematics

University of Glasgow

Learning, Teaching and Scholarship in Maths & Statistics

28th November 2023



What is a Large Language Model?

Previously on TALMO

Specifically about mathematics...

Policies

How to use LLMs

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”.

Seeing probabilities

TALMO (Teaching And Learning Mathematics Online) is a web-based system designed to help teachers and students learn and teach mathematics online. It is an interactive platform that allows teachers to create interactive lessons, assign homework, and track student progress. It also allows students to access online tutorials, practice problems, and view their grades. TALMO is designed to be used in a variety of settings, including traditional classrooms, online classes, and homeschooling.



How likely was that token?

TALMO (Teaching And Learning Mathematics Online) is a web-based system designed to help teachers and students learn and teach mathematics online. It is an interactive platform that allows users to create interactive lessons, assign homework, and track student progress. It also allows students to access online tutorials, solve problems, and view their grades. TALMO is designed to be used in a variety of settings, including traditional classrooms, online learning, and homeschooling.

web = 58.98%

free = 7.04%

computer = 5.62%

\n = 3.95%

software = 2.98%

Total: -0.53 logprob on 1 tokens
(78.57% probability covered in top 5 logits)

Why did it pick the “unlikely” token?

TALMO (Teaching And Learning Mathematics Online) is a web-based system designed to help teachers and students learn and teach

mat

to c

prog

prob

varie

and

platform = 37.55%

learning = 13.69%

mathematics = 11.26%

educational = 7.46%

system = 4.69%

Total: -3.06 logprob on 1 tokens
(74.65% probability covered in top 5 logits)

orm that allows teachers

rk, and track student

online tutorials, practice

designed to be used in a

rooms, online classes,

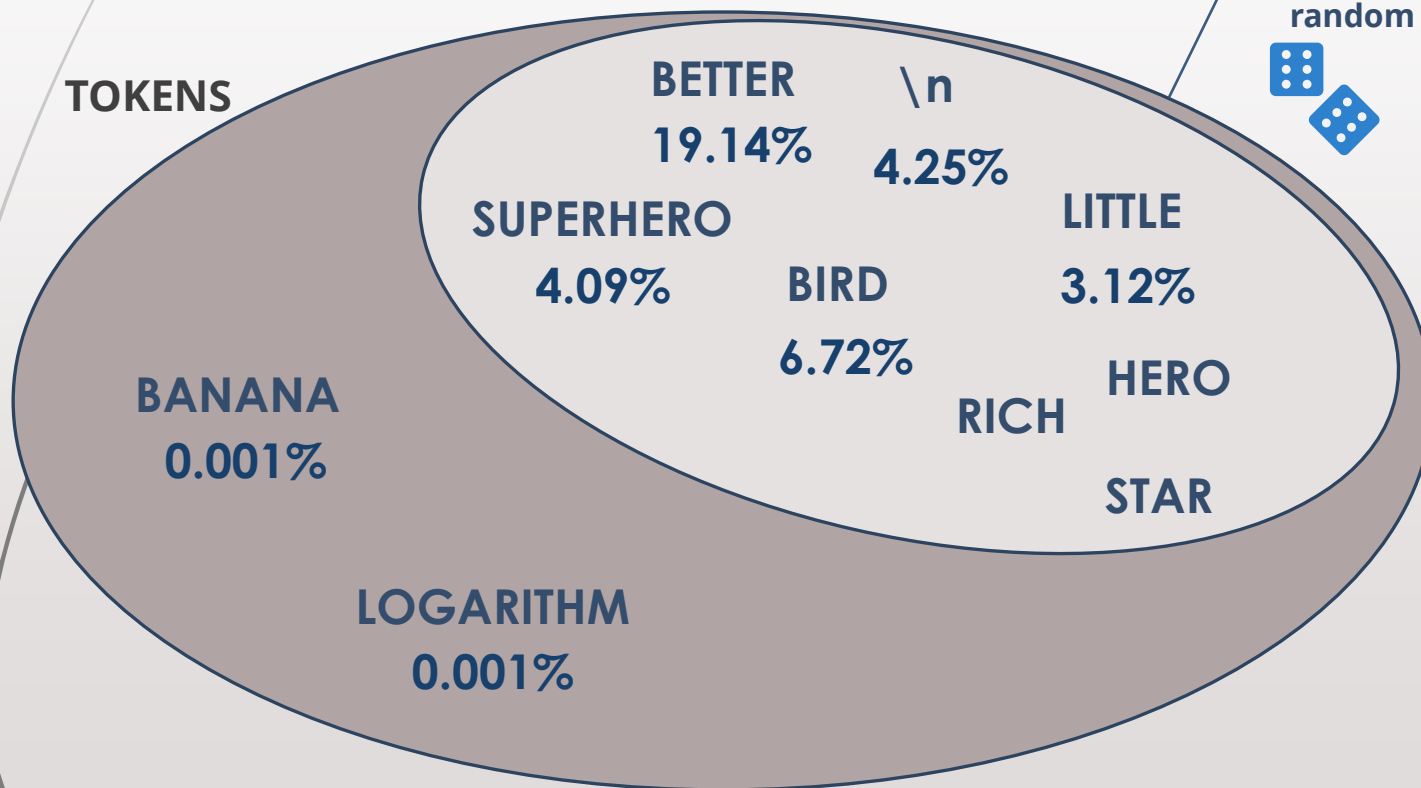
Nucleus/"Top P" sampling

PROMPT

I wish that I could be a

bird

TOKENS



Models have several parameters

Mode
☰ Complete ▾

Model
text-davinci-003 ▾

Temperature 0.69
▬ ○

Maximum length 100
○

Controls diversity via nucleus sampling: 0.5 means half of all likelihood-weighted options are considered.

Stop sequences
Enter sequence and press Tab
▬

Top P 0.7
▬ ○

Frequency penalty 0
○

Presence penalty 0
○

Best of 1
○

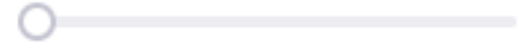
Controls randomness: Lowering results in less random completions. As the temperature approaches zero, the model will become deterministic and repetitive.

Different parameters can produce different results

TALMO (Teaching And Learning Mathematics Online) is a web-based platform designed to help teachers and students learn mathematics online. It provides a variety of tools and resources to help teachers create and deliver engaging lessons, and students to learn and practice mathematics. The platform includes interactive activities, videos, quizzes, and other resources to help students understand and apply mathematics concepts. It also provides teachers with tools to track student progress and provide feedback.

TALMO (Teaching And Learning Mathematics Online) is a software platform designed to help teachers and students learn mathematics. It is a comprehensive online system that combines multimedia teaching, assessment, and collaboration tools. The platform includes interactive lessons, quizzes, simulations, and collaborative activities that can be used to help students understand mathematics concepts. It also includes tools for creating custom lessons and activities, tracking student progress, and sharing resources.

Temperature 0



Temperature 2



What can an LLM 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 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.

LLMs tomorrow:

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

Do not focus on current generative AI. Think ahead!

Large language models and mathematics



A few technical points: an LLM...

- Can process and write symbols and notation.
- Can read and write LaTeX code.
- Can read/write/debug code for Matlab, GAP,...
- Can query WolframAlpha and interpret the results.
- Can read and process graphical data.

This is where we are:

ChatGPT 4 ▾



You

(b) [3 marks] Using logarithmic differentiation, show that if $y = y(x)$ is defined by

$$y(x) = \frac{(2x + 1)^{1/2} \exp(x)}{(3x + 1)^{1/3}}$$

then $y' = dy/dx$ can be written in the form

$$y' = f(x)y$$

and determine the function $f(x)$.

Solve (no code)



ChatGPT

To solve this problem using logarithmic differentiation, we'll take the following steps:

1. Take the natural logarithm of both sides of the equation $y(x) = \frac{(2x+1)^{1/2} \exp(x)}{(3x+1)^{1/3}}$.
2. Differentiate both sides with respect to x using the properties of logarithms to simplify the differentiation process.

This is where we are:

ChatGPT 4 ▾

This gives us the function $f(x)$ as required.



You

Give me LaTeX code for this solution



ChatGPT

Certainly! Here's the LaTeX code for the solution to the problem using logarithmic differentiation:

latex

Copy code

```
\documentclass{article}
```

```
\usepackage{amsmath}
```

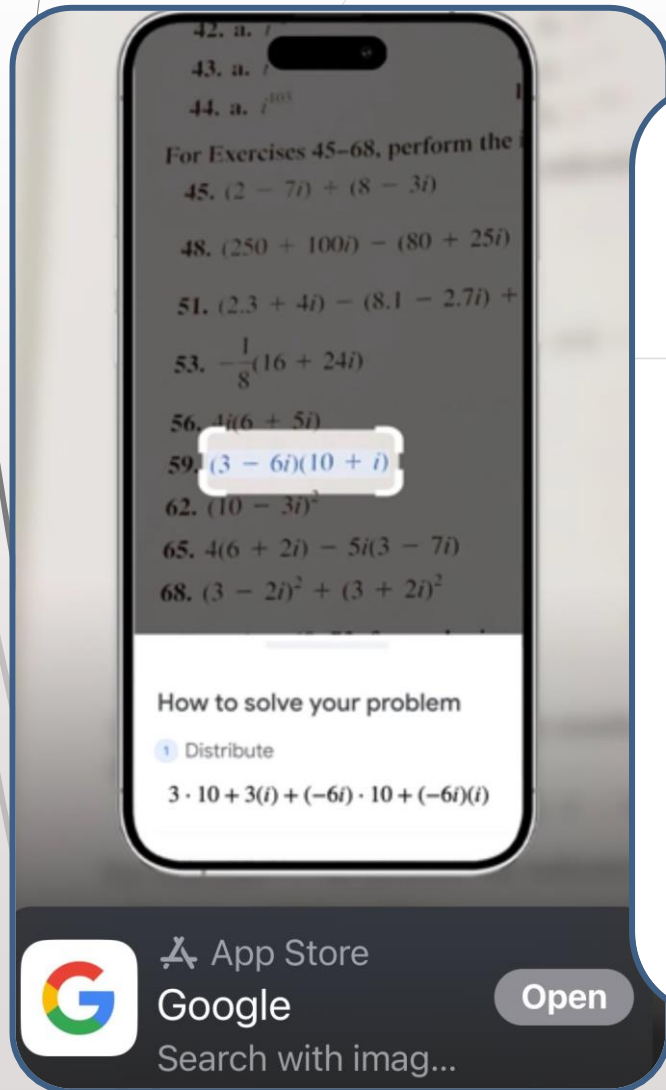
```
\begin{document}
```

```
To find the derivative of the function  $y(x) = \frac{(2x + 1)^{1/2}}{\exp(x)}$ 
```



```
Step 1: Take the natural logarithm of both sides:
```

This is where we are:



Google solve $e^{2x} = \frac{(e^x - 6)}{(e^{4x} - 2)}$

Images Videos News Books Maps Flights Finance

About 3 results (0.39 seconds)

Your maths problem

$$e^{2x} = \frac{(e^x - 6)}{(e^{4x} - 2)}$$

Find solutions on the web

Your maths problem will be shared with maths providers and processed according to their privacy policy. Your data will be shared.

Mathway Symbolab Mathforyou CameraMath

This is where we are:

- Most advanced models have AI-enhanced voice recognition.
- Models can see and interpret (and reproduce) audio, pictures and videos.
- Models are ubiquitous in websites and apps. (e.g. Microsoft 365, Grammarly)
- Models exist in a myriad of customised wrappers that significantly change their style and abilities.
- Statements like «LLMs can't do ##» should be read as «I could not find/make an LLM do ##».

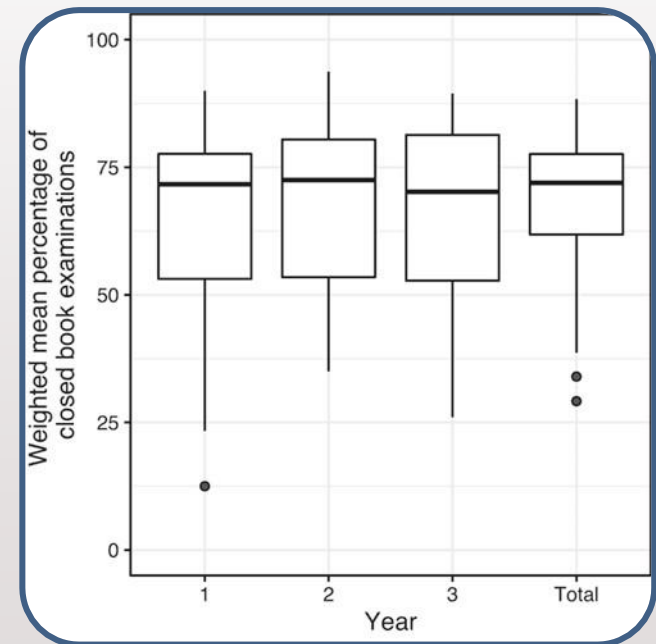
My experience

I tested a set of undergraduate mathematics questions in GPT3.5, GPT-4 and Google Bard repeatedly since January 2023.

- **Good:** calculus and equations in general. Most basic mathematics.
- **Inconsistent:** linear algebra, trigonometry.
- **Bad:** Algebra, group theory, category theory, abstract/proof based.
- It can be very inconsistent.
- It spontaneously offers structured answers, and improves further when this is explicitly encouraged.
- It defaults to standard/well-known proof structures, even when it should not.

On malpractice, we are the lucky ones

- Tools, or software, able to solve several mathematics exercises has been around for decades.
- The gold standard in mathematics assessment method are still closed book examinations.
- Mathematics learning objectives are normally broad and test students critical and foundational skills.



Iannone, Simpson (2021), [How we assess mathematics degrees: the summative assessment diet a decade on.](#)



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.**
- **Others were adapted.**
- **Others remained identical, and are assessed without access to calculators: we create controlled conditions to ensure integrity**
- **Their usage is otherwise assumed, even implicit.**

Work on learning objectives

Learning objectives	 Calculators	Large language models 
Remove Reduce Change	Compute logarithms using tables Hand-drawn geometric constructions Manual arithmetic → Less emphasis	Memory-based assessment Emphasis on writing style and spelling Factual recall → Comprehension and critical analysis
Adapt Integrate Fortify	Solve polynomial equations Compute expressions → Introduce more complex, calculator-proof exercises Introduce exercises that require competent calculator usage	Traditional essays and content creation Creative writing → Co-creation with LLMs
Keep (assess in controlled conditions)	Fundamental arithmetic skills Calculus → Non-calculator papers	Critical thinking and argument development Problem solving → In-person assessments

Safeguard human knowledge

It's not reliable in any way. The last thing you want to deal with is not only a possible misconception on your part about a topic but having to figure out chat gpt's misconceptions too. You are effectively making it 2 times worse for yourself

Student on /r/learnmath, Reddit, 2023

- Efficient, safe use of AI is only possible when the human is already skilled on the subject. The model's factual accuracy or knowledge cannot be trusted.
- LLMs get things **wrong**.
- Students need a solid background to be able to critically evaluate the output of a model.
- Education remains as important as ever.

An example I like

- *“If we train a model on text from a medieval library, then it won’t be able to figure out that the Earth is not at the center of the Solar system” (me, Spring 2023)*

MonadGPT

New Chat

Outer Space Materials

What is outer space made of?

- The Heavens and Earth doe not contain them, but they contain all things: The Sunne is a body of Fire, and the Moon a body of Earth, and the Stars are bodies of Water, or Aire, which all these have in them, and therefore they beare the names of the four Elements. And when they come to a higher degree of heat, they change into Fire, which is the most excellent and perfect of all Bodies; for it hath no part, nor figure, nor quantity, nor weight, nor bulk, nor any thing that can be discerned by the senses, but it is a pure Spirit, and as a Body of Light, and being so, it is invisible, and yet it is more substantial then all other Bodies, because it is the first and last of all Creatures, and it is the foundation and substance of all things.

Russell Group Universities

- Universities will support students and staff to become AI-literate.
- Staff should be equipped to support students to use generative AI tools effectively and appropriately in their learning experience.
- Universities will adapt teaching and assessment to incorporate the ethical use of generative AI and support equal access.
- Universities will ensure academic rigour and integrity is upheld.
- Universities will work collaboratively to share best practice as the technology and its application in education evolves.

Qualities of a good policy

- **Precise:** avoids vague statements and uncertainty.
- **Explicit:** endorses and encourages responsible AI usage.
- **Honest:** transparent about capabilities and limitations of AI.
- **Fair:** ensures equal scrutiny and safeguards.
- **Inclusive:** ensures equal access and opportunities.
- **Robust:** academic integrity is still safeguarded.
- **Adaptable:** considers the rapid development of AI.
- **Human-centric:** aims to preserve human learning and skills.

This also applies locally, to course policies!

A word on detection...

- Detectors have significant false positive rates.
- Detectors are vulnerable to basic attacks.
- Detectors are not falsifiable.

Further...

- Detectors cannot distinguish malpractice from good usage.
- They could find a place in a “full ban” policy, but as soon as any usage of AI is permitted, they become detrimental.
- There is a strong business incentive to sell (flawed) AI detection technology to institutions.

Avoid AI detection, and speak up against its implementation.

University of Stanford

Absent a clear statement from a course instructor, use of or consultation with generative AI shall be treated analogously to assistance from another person. In particular, using generative AI tools to substantially complete an assignment or exam (e.g. by entering exam or assignment questions) is not permitted. Students should acknowledge the use of generative AI (other than incidental use) and default to disclosing such assistance when in doubt.

Individual course instructors are free to set their own policies regulating the use of generative AI tools in their courses, including allowing or disallowing some or all uses of such tools. Course instructors should set such policies in their course syllabi and clearly communicate such policies to students. Students who are unsure of policies regarding generative AI tools are encouraged to ask their instructors for clarification.

University of Boston

Students shall

1. Give credit to AI tools whenever used, even if only to generate ideas rather than usable text or illustrations.
2. When using AI tools on assignments, add an appendix showing (a) the entire exchange, highlighting the most relevant sections; (b) a description of precisely which AI tools were used (e.g. ChatGPT private subscription version or DALL-E free version), (c) an explanation of how the AI tools were used (e.g. to generate ideas, turns of phrase, elements of text, long stretches of text, lines of argument, pieces of evidence, maps of conceptual territory, illustrations of key concepts, etc.); (d) an account of why AI tools were used (e.g. to save time, to surmount writer's block, to stimulate thinking, to handle mounting stress, to clarify prose, to translate text, to experiment for fun, etc.).
3. Not use AI tools during in-class examinations, or assignments, unless explicitly permitted and instructed.
4. Employ AI detection tools and originality checks prior to submission, ensuring that their submitted work is not mistakenly flagged.
5. Use AI tools wisely and intelligently, aiming to deepen understanding of subject matter and to support learning.

University of Glasgow

What is the University of Glasgow's position on use of AI?

The University of Glasgow (with Russell Group peers) believes artificial intelligence (AI) tools are potentially transformative as well as disruptive. They increasingly feature in academic and professional workplaces. Consequently, **rather than seek to prohibit students' use of these tools, we want to support students in learning how to use them effectively, ethically, critically, and transparently.** Our students will graduate into an AI-augmented world. The university has a responsibility to prepare students for this world, providing space to experiment with, and understand the potential of, AI in an ethical way.

University of Glasgow

A summary of some of the common ways that students can appropriately use Generative AI are listed below:

- Answering questions where answers are based on material which can be found on the internet.
- Drafting ideas and planning or structuring written materials.
- Generating ideas for graphics, images and visuals.
- Gathering information for subsequent critical review and analysis.
- Helping to improve grammar and writing structure – especially helpful if English is a second language.
- Experimenting with different writing styles.
- Getting consolidated explanations.
- Debugging code.
- Getting over writer's block.

University of Glasgow

Update examinations

In-person (invigilated) examinations are not the ideal solution to overcoming the challenges that AI tools pose for assessment. From extensive research, we know that that invigilated examinations are not conducive to meaningful or inclusive assessment (see [Learning Through Assessment](#) p. 9 & 15). Recent and ongoing development of AI tools requires us to consider these in the context of different types of examinations. While there is still a place for invigilated examinations, we must consider how to bring AI into that setting. For example, rather than getting students to engage in knowledge recall in examinations (e.g., reproducing a mathematical proof or derivation, or writing down the definition or characteristics of a certain type of behaviour) you could use AI to generate these items and then require students to critique or fix certain problems. In addition, rather than asking students to generate an article on a contemporary topic, you could generate this through AI and require students to identify areas for improvement and/or critique the way it is constructed. Alternatively, there is an opportunity to draw upon unique data sets that may have been introduced during the semester, to create examinations that require students to engage in critical analysis and apply skills during the examination.

GPT-4 can do that

Advanced Data Analysis can do it

Even GPT-3.5 can do this!

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.**

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.
- Students need to be taught why it is still important for them to learn the skills that LLMs can (or appear to) emulate.
- The ultimate goal is to encourage and allow appropriate LLM usage, while preserving the authenticity of assessment.

AI should be a copilot, never a pilot.

Learning to use 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.

A desirable endgame



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



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



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



- Robust, authentic assessment is developed and acts as an effective motivator to encourage students to learn skills, regardless of LLM 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, or using online platforms that can track input and the progress of the work*.



- 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.

How to use ChatGPT & friends

Most available models have been trained to be effective conversational chatbots (or tools based on those).

- 
- Use natural language.
 - Just talk to the language model.
 - Explain what you want to do, and how you want the output to look like.
 - Be critical: AI should be a copilot, never a pilot.
 - As AI gets better, it will get better at understanding what you want. Prompt engineering is a short-term workaround.

Be mindful of privacy, especially when using third party tools.

Thank you for your attention

Feel free to contact me:

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

I am interested in ~~experiments~~ case studies that involve students and integrate AI, or analyse its usage, or co-develop resources or assessments, etc.