

Large language models and mathematics higher education

Cesare G. Ardito



Teaching Fellow → Lecturer
Department of Mathematics
University of Manchester

TALMO (Teaching And Learning Mathematics Online)
7th June 2023



What is a Large Language Model?

Large language models and
mathematics

What should we do?

Opportunities

Homework

Why is Calvin wrong?

GIVEN THE PACE OF
TECHNOLOGY, I PROPOSE
WE LEAVE MATH TO THE
MACHINES AND GO PLAY
OUTSIDE.



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	platform = 37.55%	orm that allows teachers
to c	learning = 13.69%	rk, and track student
prog	mathematics = 11.26%	online tutorials, practice
prob	educational = 7.46%	esigned to be used in a
varie	system = 4.69%	rooms, online classes,
and		

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

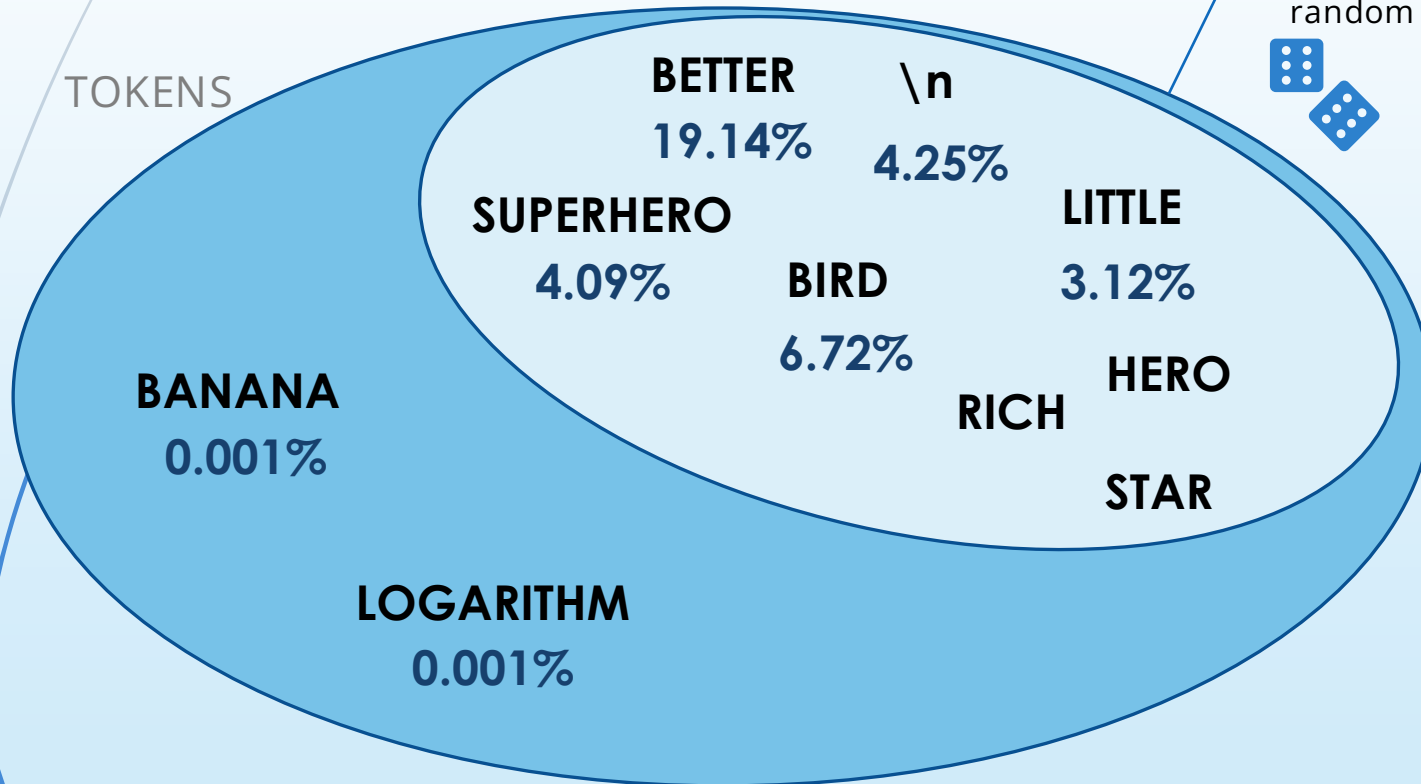
Nucleus/"Top P" sampling

PROMPT

I wish that I could be a

bird

TOKENS



Models have several parameters

Mode

Model

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.

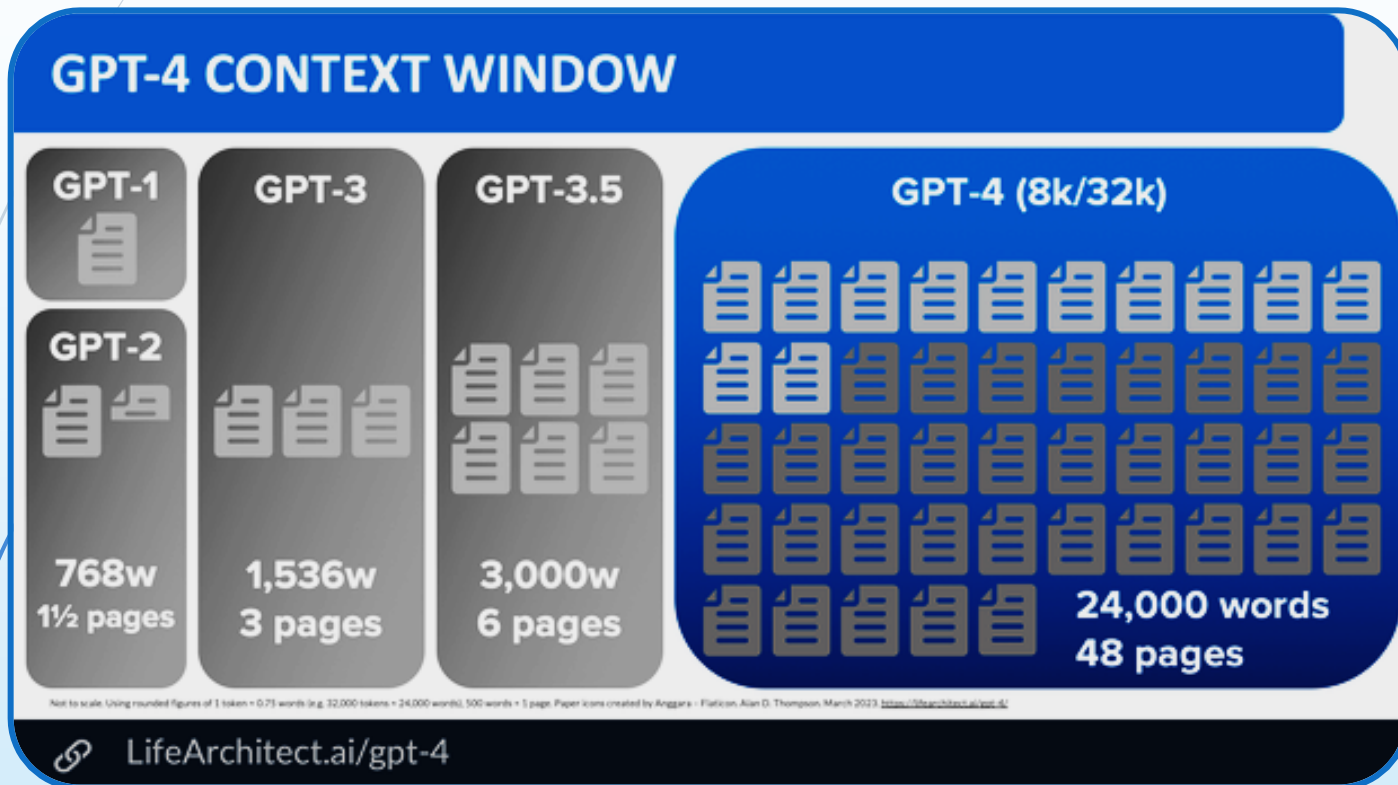
Temperature 0



Temperature 2



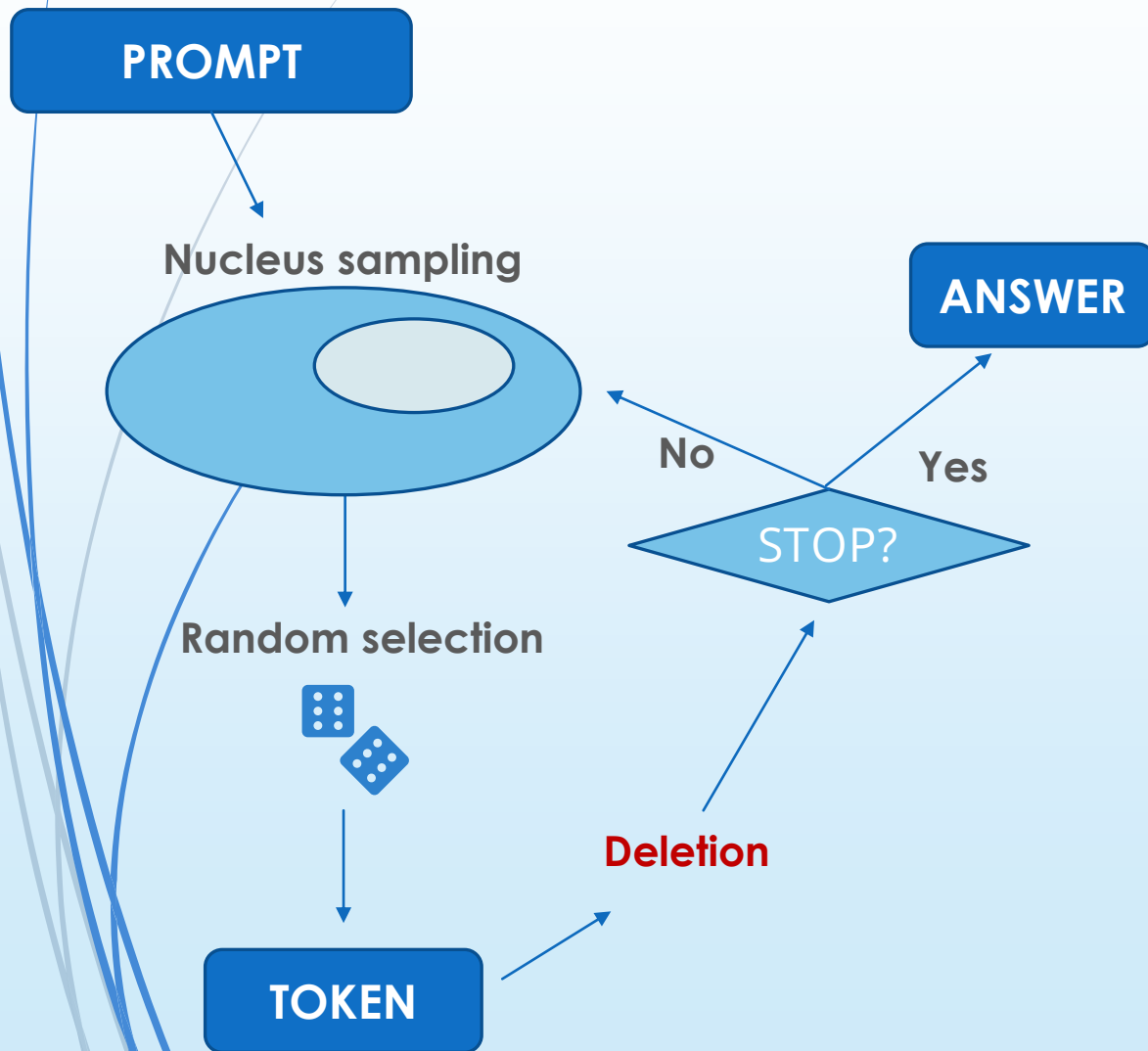
Context window



Close to the memory capacity of the Commodore 64



To summarise:



Initial training
+
Reinforcement learning
from human feedback
+
Parameters

determine the weights for
each token



GPT-4 explains training with
cake metaphors

One thing that ChatGPT cannot do

⚡ Model: Default (GPT-3.5)



Think of a number between 1 and 10, and I will try to guess it. Ready?



Sure, I'm ready! I've thought of a number between 1 and 10. Go ahead and make your first guess.

< 1/2 >



Is it 9



Congratulations! You got it right!

< 2/2 >



Is it 9?



No, it's not 9 either. Keep guessing!

However, Bing AI can play “guess”



tuneworm (Joaquin Castellano)

@StudentInfosec

After interpreting the message, Bing runs an internal command called `#inner_monologue`. In here it decides on the language for the message, and how to generate its response — whether it’s necessary to perform a web search, or if it should provide product ads

```
#inner_monologue
```

```
Deciding on language:
```

```
`English` detected from the  
user message. Using the same  
language for my reply and my  
suggested user responses.
```

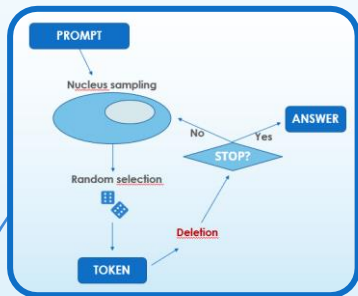
```
Deciding on search need:
```

```
`No`. The response cannot be
```

...because Bing AI is not just a LLM

PROMPT

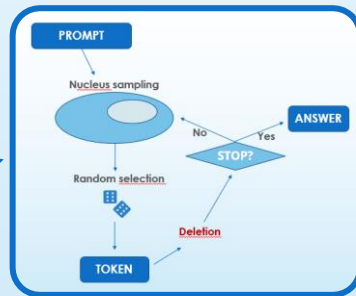
#inner_monologue



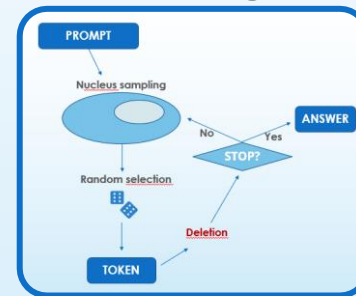
No

Yes

#inner_monologue



#message



ANSWER

PLUGIN

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.

5,000+ apps

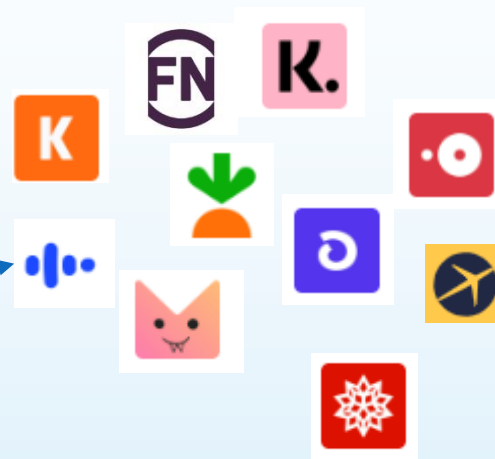


Zapier




GPT-4

Plugins



API



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



ChatGPT



Copilot

Stanford
Alpaca



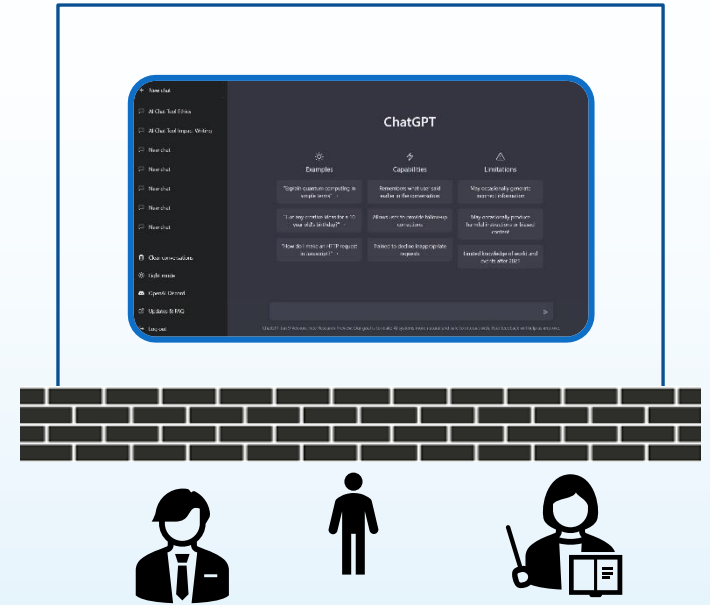
Bard



Meta AI

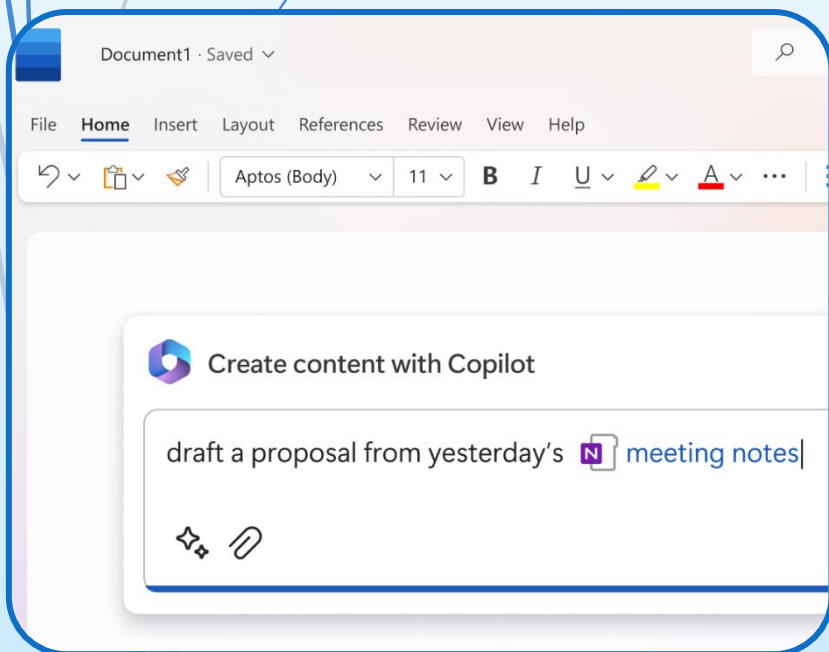
LLMs today:

- 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

be my eyes OpenAI

User Answer question I.I.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.

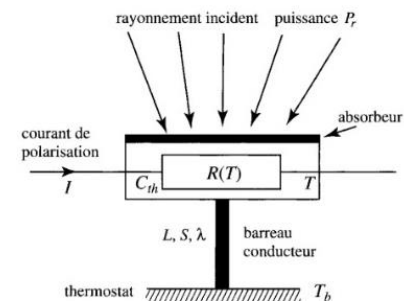


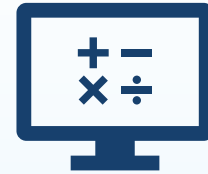
Figure 1 - Schéma d'un bolomètre

LLMs tomorrow:

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

Do not focus on current generative AI. Think ahead!

Large language models and mathematics



A few technical points: a 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.**

NEW

SOON

A few studies

January 2023:

- “ChatGPT’s mathematical abilities are significantly below those of an average mathematics graduate student”.

(Frieder, S., Pinchetti, L., Griffiths, R., Salvatori, T., Lukasiewicz, T., Petersen, P., Chevalier, A., & Berner, J. (2023). Mathematical Capabilities of ChatGPT, <https://arxiv.org/abs/2301.13867>)

February 2023:

- “We found that ChatGPT's performance changes dramatically based on the requirement to show its work, failing 20% of the time when it provides work compared with 84% when it does not”.

(Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). An Independent Evaluation of ChatGPT on Mathematical Word Problems (MWP), <https://arxiv.org/abs/2302.13814>)

A few studies

April 2023:

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)

(OpenAI (2023), GPT-4 Technical Report, <https://arxiv.org/abs/2303.08774>)

May 2023:

- “Our findings revealed that all three chatbots demonstrated to some extent an ability to understand and process math and logic problems, with some exceptions and limitations. These models can be used to solve basic mathematics and logic problems”.

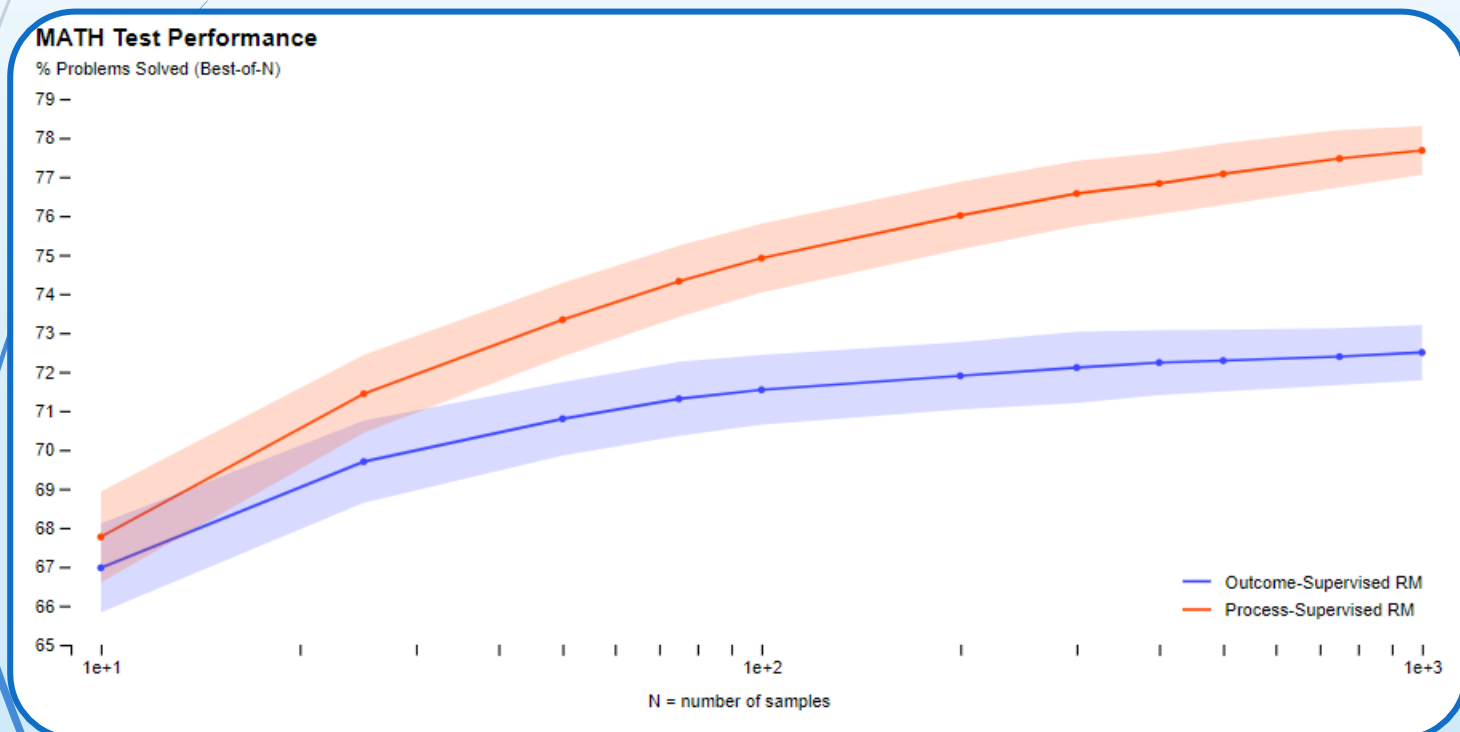
(Plevris, V., Papazafeiropoulos, G., & Rios, A. J. (2023). Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard,

<https://arxiv.org/abs/2305.18618>)

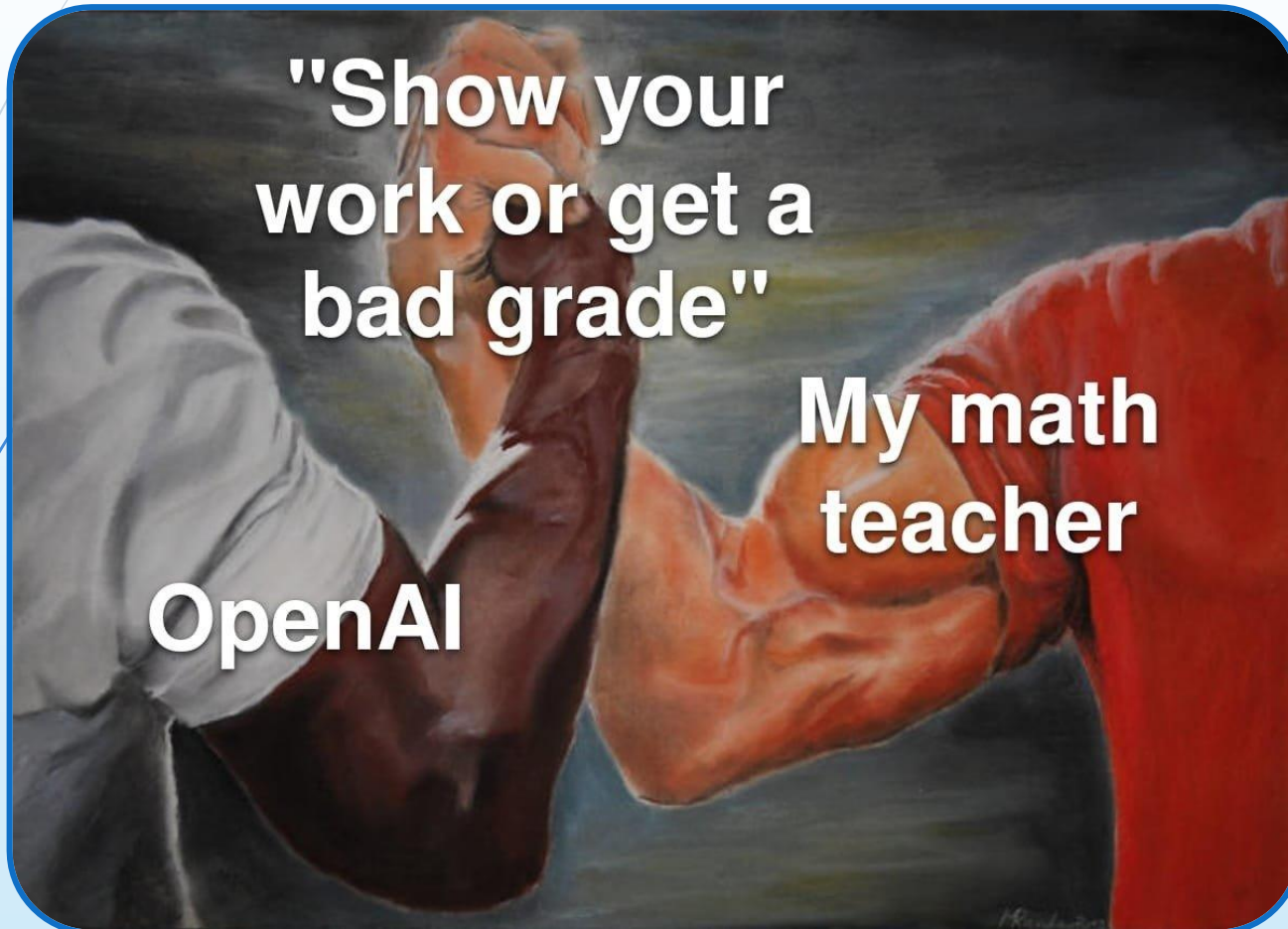
A very recent study

“We’ve trained a model to achieve a new state-of-the-art in mathematical problem solving by rewarding each correct step of reasoning (“process supervision”) instead of simply rewarding the correct final answer (“outcome supervision”).”

(OpenAI (2023), [Improving mathematical reasoning with process supervision](https://arxiv.org/abs/2305.20050), <https://arxiv.org/abs/2305.20050>)



A very recent study



AI Notkilleveryoneism Memes, Twitter

My experience

I tested a set of undergraduate mathematics questions in ChatGPT, GPT-4, Bing AI and Google Bard repeatedly during the last few months.

- It does not yet perform well at advanced (Y3+) mathematics topics.
- It can pass some Y1 and Y2 assessments, but several answers are often incorrect or incomplete.
- 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.
- It is improving a lot: GPT-4 excels at some tasks that it could not do even just a month ago.

My experience

ChatGPT

Jan 30 version update

Here's what's new

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

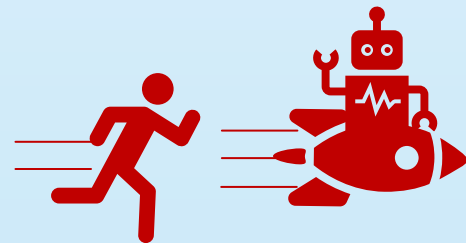
Sounds good!

"I have quickly reran the same prompts I used in my talk, and it seems that ChatGPT is now significantly better than before at solving maths problems. Most of the mistakes it made on the slides have disappeared (e.g. it computes squares mod 9 correctly). It has also started using some sort of $$ environment for formulae.

Fresh evidence to validate the last 20 minutes of the talk, I suppose!"

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.



Why is Calvin wrong?

GIVEN THE PACE OF
TECHNOLOGY, I PROPOSE
WE LEAVE MATH TO THE
MACHINES AND GO PLAY
OUTSIDE.



Because of a fundamental issue:

- We cannot trust LLM output to be accurate/correct/true.
- At the moment, LLMs are very inconsistent in mathematics.



Let K be a field. Let b be an element of K . If $xb=x$ for all x in K . then $b=1$.



Assume that $b \in K$ is such that $xb = x$ for all $x \in K$.

We can choose $x = 0$ in this equation, which gives $0b = 0$.

Since b cannot be 0, as it is an element of a field, we can divide both sides by b , giving $0 = 0$.

This means that $b = 1$.

- As educators, it is crucial we preserve our students' critical thinking (and human critical thinking in general).
- This is especially important in mathematics.



Let K be a field. Let b be an element of K . If $xb=x$ for all x in K , then $b=1$.



If for all x in K , $xb = x$, then taking $x = 1$ (the multiplicative identity in the field K), we have $1b = 1$. But $1b = b$, hence $b = 1$, which completes the proof.



(full disclosure: GPT-4 can now do this well – it is now, in general, hard to find an "easy" exercise that GPT-4 gets consistently wrong)

Are students using LLMs?

Yes, but:

I recommend everyone to not use ChatGPT for math since it has given me nonsensical arguments or claimed things to be true which obviously are not.

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

It's really hit or miss. I've fed ChatGPT problems from my first year physics textbook and it often calculates the correct answer. Then, like you say, it makes some glaring and obvious mistake.

(Mathematics) Chatgpt is so wrong but also stubborn to accept it

Most feedback available from students suggests that using an LLM to assist in solving exercises in mathematics is ineffective.

Are students using LLMs?

However:

TLDR: ChatGPT appears to be able to figure out in the midst of a response if it is trying to prove a false statement and is capable of giving correct proofs from that point onward.

A majority of the respondents believe that chatbots and AI language tools make them more efficient as students and argue that such tools improve their academic writing and overall language skills. Virtually all the responding students are familiar with ChatGPT, the majority use the tool, and 35 percent use the chatbot regularly.

Many students perceive chatbots as a mentor or teacher that they can ask questions or get help from, for example, with explanations of concepts and summaries of ideas. The dominant attitude is that chatbots should be used as an aid, not replace students' own critical thinking. Or as one student put it: "You should be able to do the same things as the AI, but it should help you do it. You should not use a calculator if you don't know what the plus sign on it does".

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

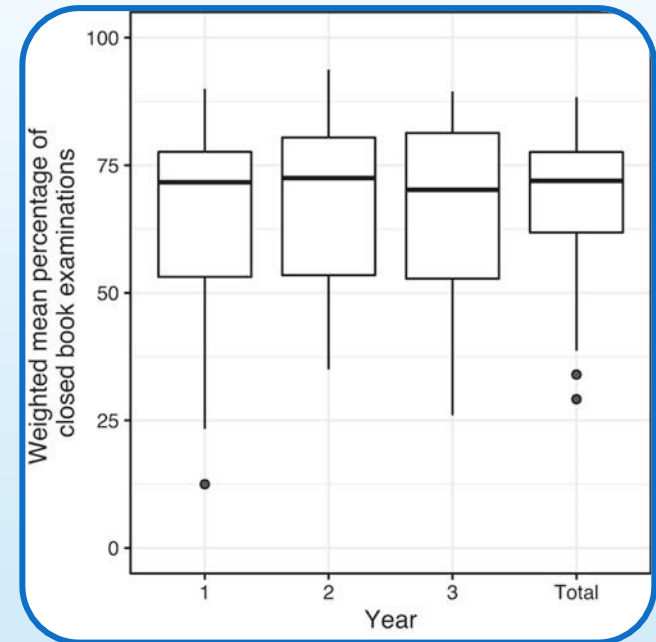
Teaching students about LLMs is not optional

- 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.

My suggestion: a departmental training module.

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.
- Mathematical innovation involves creativity and genuine novelty, which AI visibly struggles to emulate.



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

Push back against detection

- Detectors are currently unreliable and fundamentally unreliable*.
- 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.
- **There is a strong business incentive to sell (flawed) AI detection technology to institutions. Do not fall for it: hold the line!**

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

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, 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.

That is all!

Opportunities

- LLMs can be used to facilitate marking.
- LLMs can facilitate several admin tasks, freeing up time for more important work.
- LLMs can help making documents and resources more accessible.
- Integrated LLM will change how we use most tools.

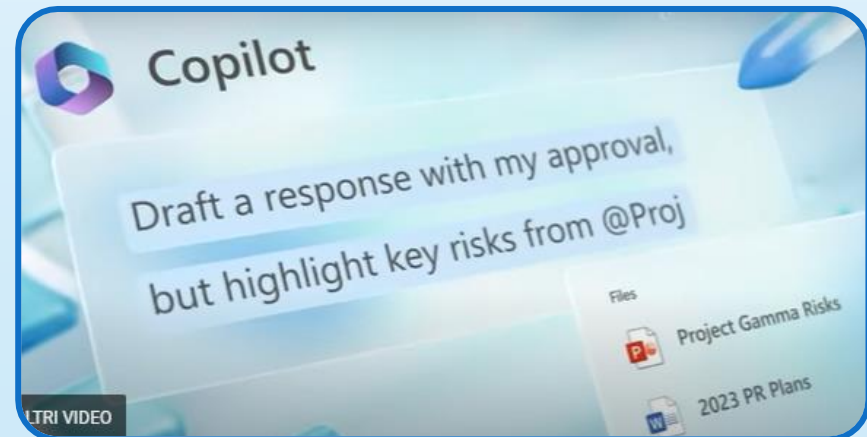
AI should be a copilot, never a pilot

Marking

- **Productivity boost:** auto-adding scores, spreadsheets, etc.
- **First-pass review** of responses, highlight potential errors, suggest corrections.
- **Standardization:** reduce bias and improve consistency.
- **Feedback generation:** based on the marker's input.
- **Small context window:** current LLMs cannot process a project or a long essay in its entirety.
- There is AI marking software, but I am not aware of commercial software currently using LLM-based AI.
- ! Be careful: students tell us that they value and expect human teachers' feedback.

Admin

- Auto-write emails, letters from prompts.
- Write code to automate processes.
- Process data into spreadsheets, forms, etc.
- Generate slides for talks.
- Microsoft Copilot and Bard in Google Docs are the most promising upcoming technologies.

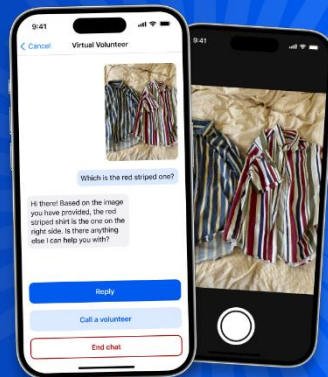


Accessibility

- Automatic conversion of resources into accessible ones.
- Relaxation of accessibility requirements to due better helper tools (see BeMyEyes).
- Text-to-speech and viceversa. Screen readers.
- Instant contextual translation.

Introducing Virtual Volunteer™

AI powered Visual
Assistant



Introducing ChatGPT and Whisper APIs

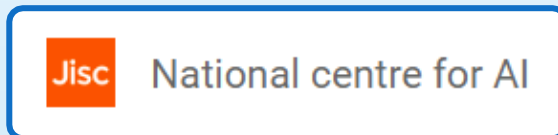
Developers can now integrate ChatGPT and Whisper models into their apps and products through our API.

Homework

- Engage with ChatGPT, or Bing AI, or Google Bard.



- Read JISC's generative AI summary post.



Michael Webb.

- Make your own toy LLM and play with it.

Seriously, do it!



[Code: Andrej Karpathy, OpenAI.](#)

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, “A generative AI primer”, JISC. <https://nationalcentreforai.jiscinvolve.org/wp/2023/05/11/generative-ai-primer/> .
- 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](#).