KING'S College LONDON

Understanding the latest advances in AI tools: How do Large Language Models work?

Andrew Woodard
BALEAP PIM, Durham
30th June 2023

Outline

The basic rationale here	
Instances of LLMs revealing how they work	**Non-technical technical tangents**
Two fallacies	Fine-tuning (aka ‘RLHF’)
LLMs don’t understand the world	Thought vectors (aka ‘embeddings’)
More examples of LLMs not understanding how the world works	
Reference hallucinations	The probability (temperature of 0.8)
Other instances of 0.8 temperature?	

Disclaimer(s)

I am not an expert on these matters. I'm an EAP teacher!

And this is a work in progress (I still don't understand how LLMs work to my own satisfaction)—so 'filling the gap' (as per my abstract) between simplified the descriptions and the computer science literature was rather ambitious.

If anything I wrote off Machine Learning some time ago (it's not how the mind works, and Google Translate couldn't translate 'The apple ate John' into French, at least before it embrace Deep Learning, pre-2016).

But then I tried ChatGPT…

…and I was very impressed

…and immediately realised it affected me as an EAP Teacher

…and I just want to know how something non-human can produce such human-like responses.

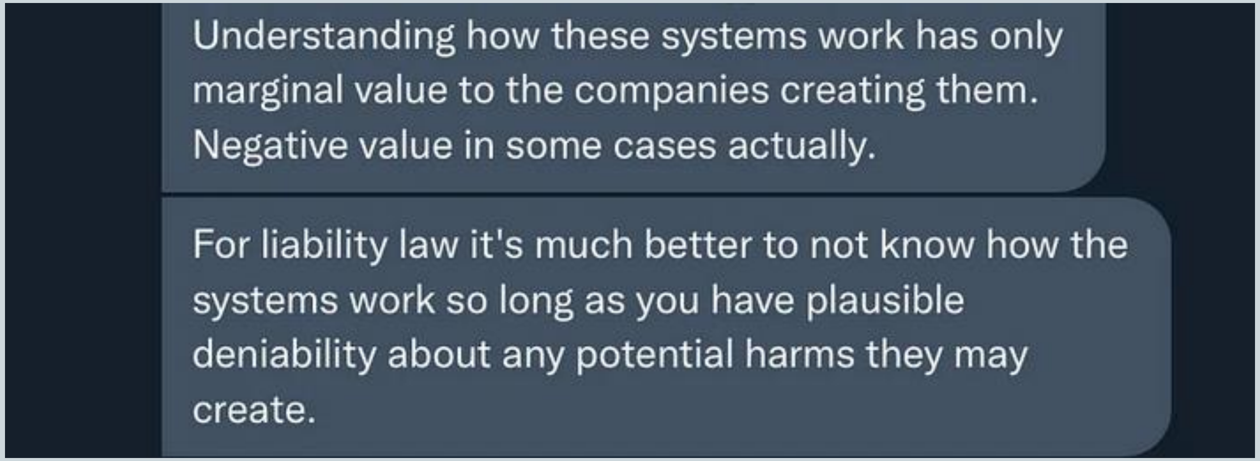
So, the next 20 minutes will be a slight intermission in terms of the pedagogical debates we're mainly interested in.

There is a degree of non-transparency

I still don’t understand how LLMs work to my own satisfaction—but then I’m not sure anybody does, especially if you don’t work for OpenAI.

> ‘OpenAI is keeping their algorithms close to their chest…they are not being transparent about the hierarchy, structure and training protocols of GPT’ (Will Jones, Director of Research, Hull University, Data Science, AI, and Modelling Centre, p.c.).

Blake Lemoine, ex-Google Engineer (Marcus, 2022):


Understanding how these systems work has only marginal value to the companies creating them. Negative value in some cases actually.

For liability law it's much better to not know how the systems work so long as you have plausible deniability about any potential harms they may create.

This non-transparency, incidentally, makes them unlike calculators (whose outputs can be fully explicated).

The basic rationale here

If we can understand how these systems work, it might help us somehow.

Isn’t that what we generally do with new technology?

Problem here: Even if we do know, it might be that this particular innovation is similar to the capacity of an anonymous human producing writing on demand.

Instances of LLMs revealing how they work—two fallacies

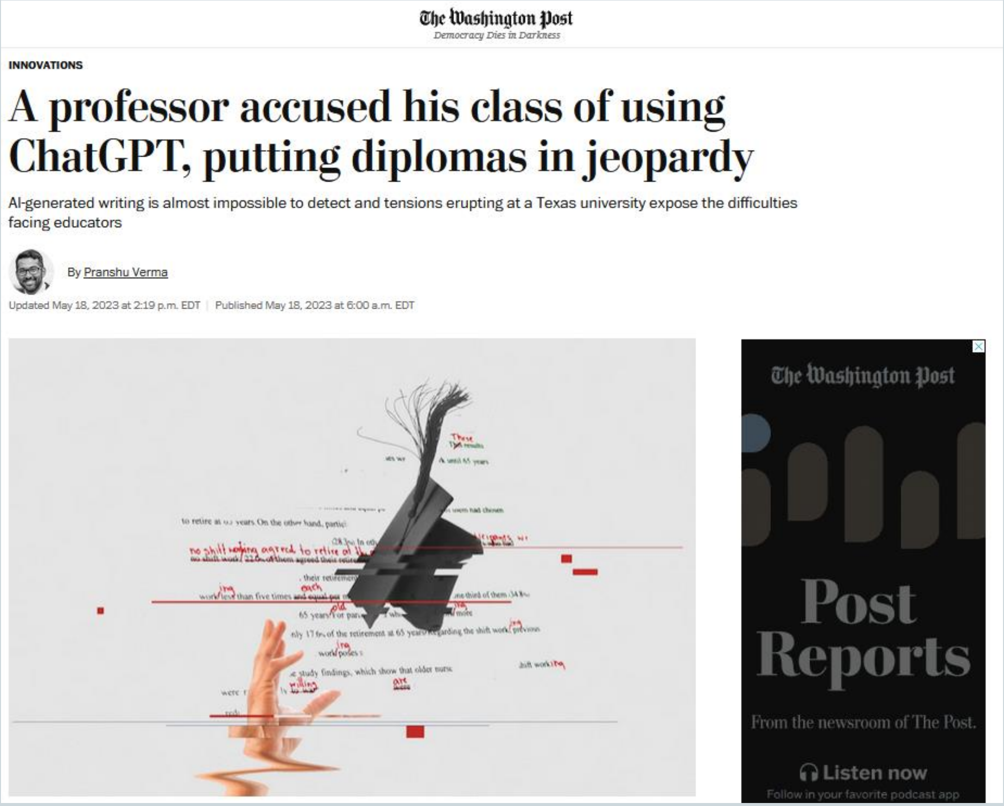
There was at least one early fallacy: the belief that ChatGPT could be used to detect whether it had written an essay you gave it. Mid-May: Animal science instructor at Texas A&M university reported in *Washington Post* (May 19, 2023). (In that article, Eric Wang, Turnitin's vice president of AI says 'Detecting AI generated text is hard. The software searches lines of text and looks for sentences that are "too consistently average"'.)

Another fallacy:

Prompt—'Can you summarise this Youtube video about the life of a baby polar bear in the Arctic: https://www.youtube.com/watch?v=9vgnXRypc4o?'

Response: 'Sure, the video…follows a baby polar bear as she learns to survive in the Arctic. The video shows the bear hunting for seals, playing with her siblings, and learning how to swim…'

LLMs can't watch videos and instantly summarize them. Here Google Bard is just riffing off the information I've given it in the prompt (the baby bear in the video is an only child for a start).


The Washington Post
Democracy Dies in Darkness

INNOVATIONS

A professor accused his class of using ChatGPT, putting diplomas in jeopardy

AI-generated writing is almost impossible to detect and tensions erupting at a Texas university expose the difficulties facing educators

By Pranshu Verma

Updated May 18, 2023 at 2:19 p.m. EDT | Published May 18, 2023 at 6:00 a.m. EDT

Tangent—fine-tuning

Such confident untruths (produced for both those two fallacies) is confusing and this, I think, reflects how LLMs work, specifically the fine-tuning that distinguishes ChatGPT from GPT3 (which was just an autofill completion of sentences, rather than a conversational interaction).

As one writer has put it:
'Unlike many other LLMs, ChatGPT is trained using reinforcement learning from human feedback [RLHF]…To train it, OpenAI had human reviewers rank outputs generated from the initial model to create a reward function that could be used to fine-tune the model (ie. make it more conversational)' (Caldwell-Gatsos, 2022).

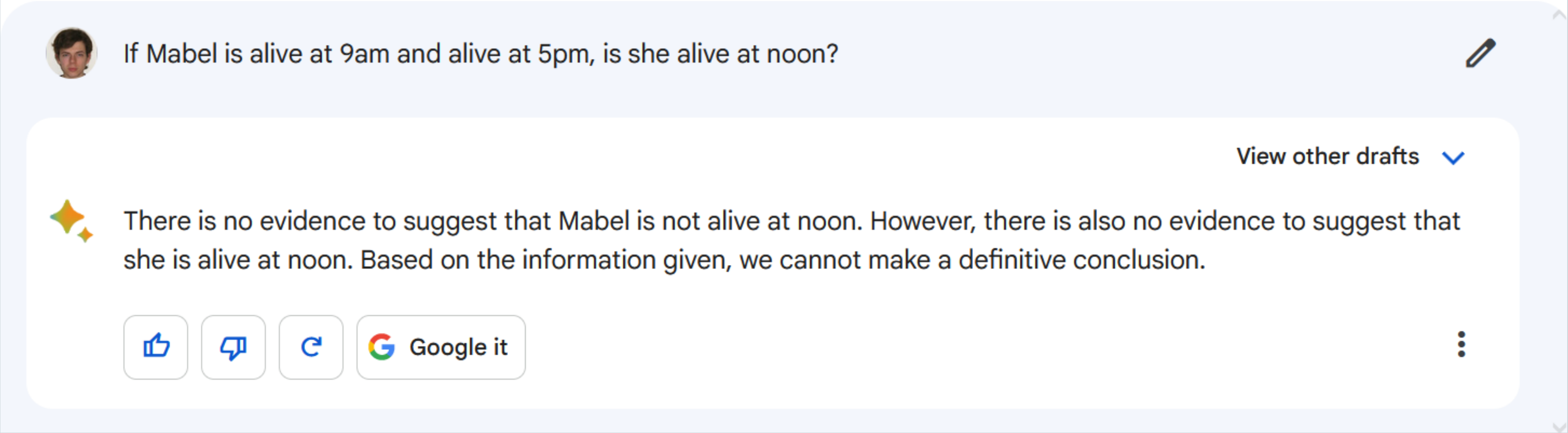
Stephen Wolfram YouTube video (c. 2:09:30, based on his 2023 article):
'Six months, a year ago…the text they produced was only so-so…Then humans rated what came out and said "That's terrible", "that's better"….And that little bit of poking seems to have had a very big effect'.

—so I think we're just seeing that *kind* of speech was rewarded. LLMs seem eager to be helpful and give a response to a prompt even when it's not something they can actually do.

LLMs don’t understand the world

Also if you understand how LLMs work—and know that they don’t understand the world—it’s easy to elicit mistakes that children wouldn’t make.

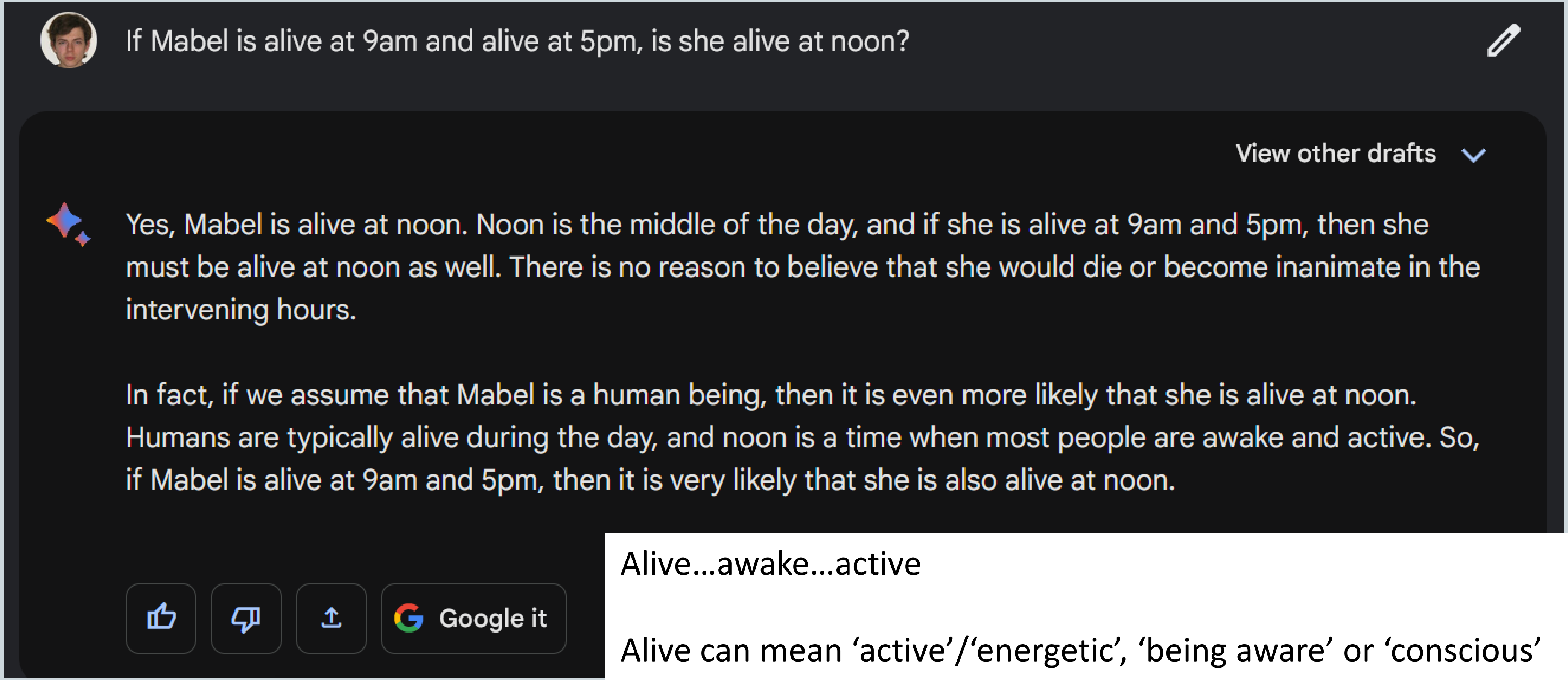
Steven Pinker (February 14, 2023):
‘This year I was terrified that that part of the lecture would be obsolete because the examples I gave would be aced by GPT. But I needn’t have worried. When I asked ChatGPT, “If Mabel was alive at 9 a.m. and 5 p.m., was she alive at noon?” it responded, “It was not specified whether Mabel was alive at noon. She’s known to be alive at 9 and 5, but there’s no information provided about her being alive at noon”’ (Powell, 2023).



Google Bard, 28th March 2023
--there *is* evidence that Mabel is alive at noon, viz. her being alive at 5pm!

It’s still not quite right

Google Bard, about 3 days ago:



Alive…awake…active

Alive can mean ‘active’/‘energetic’, ‘being aware’ or ‘conscious’ of something (which is perhaps like being ‘awake’)

Is this to do with ‘thought vectors’…?

Tangent—thought vectors

Credited to Geoff Hinton (Marcus and Davis, 2019, p.128) (technically called ‘embeddings’).

Words with similar meanings have similar vectors. For example (the real vectors are much longer lists of numbers):

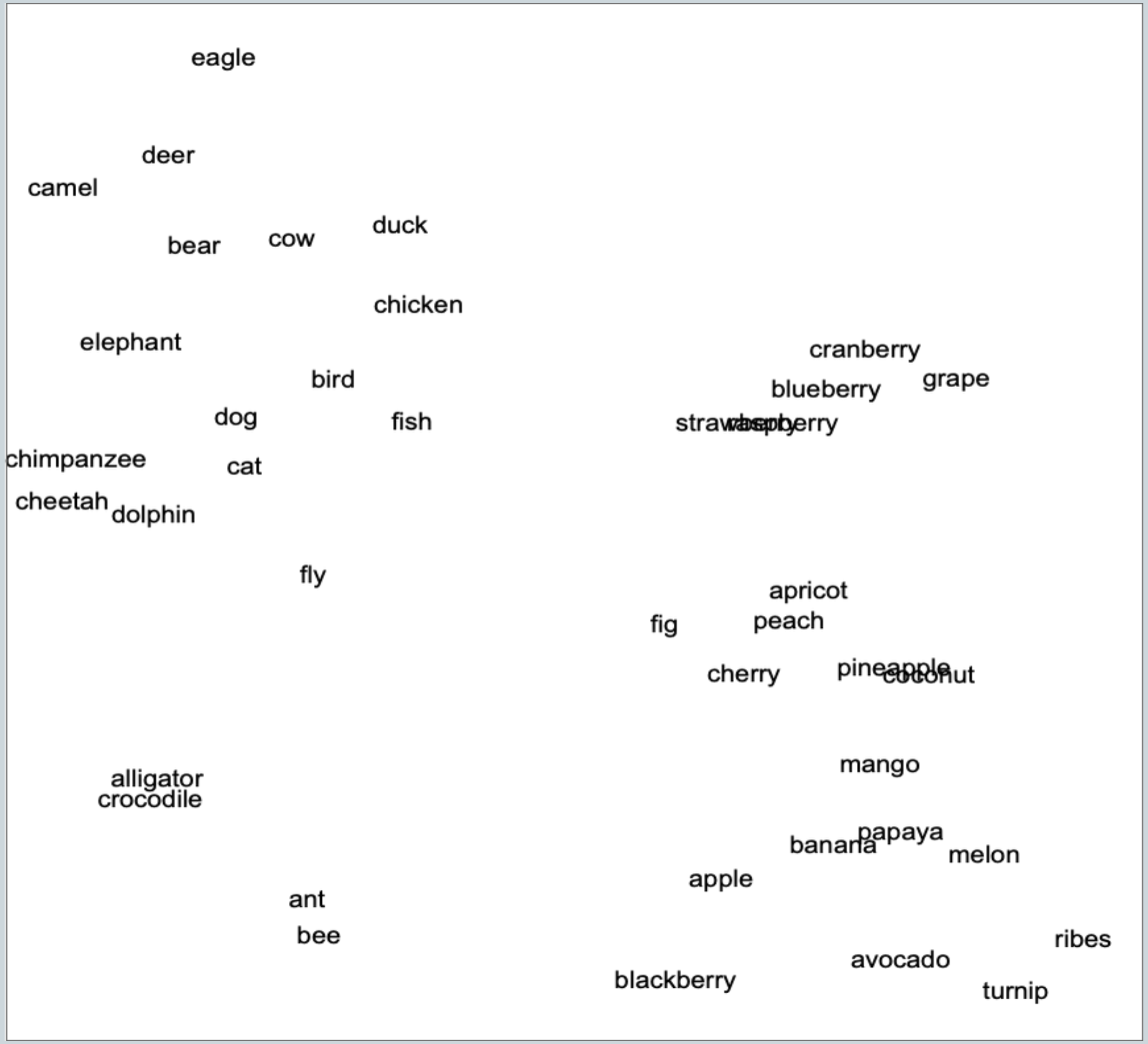
If *cat* is [0, 1, -0.3, 0.3], then *dog* might be [0, 1, -0.35, 0.25], and *train* [4, 6, 0.54, 0.87].

These numbers are based on similarities in the linguistic environments of the words. The system has seen enough sentences talking about cats and dogs in a similar way, which is to say dissimilar to the way we talk about trains. People talk about being late for trains, train tickets being expensive, train journeys being long, all different to the way we talk about pets—feeding them, taking them to the vets. LLMs pick up on this.

Famous result: In 2013 Word2Vec (system for creating embeddings developed by Ilya Sutskever, one of the founders of OpenAI) correctly guessed verbal analogies e.g. *man is to woman as king is to*____.
Add together the numbers representing *king* and *woman* and subtract the numbers representing *man*, and then look for the nearest vector: the answer was *queen*. Or as Geoff Hinton said, ‘If you take the vector for *Paris* and subtract the vector for *France* and add *Italy,* you get *Rome*…It’s quite remarkable’ (Marcus and Davis, 2019, p.129).

Tangent—thought vectors

Simplified 2D representation (from Wolfram, 2023) where the vector would be two numbers, representing (on the x-axis) broadly fruitiness (apart from turnip) and (on the y-axis) broadly animal-ness. Here we see *turnip* and *eagle* are more frequently seen in dissimilar linguistic environments.



Back to them not understanding how the world works

Showing students examples of LLMs not solving problems a child would solve (Marcus and Davis, 2023). Why was I doing this? To warn them about uncritically trusting them? Or just warning them off blind plagiarism, given the students would easily detect these mistakes?

DA

Two children, Chloe and Alexander, went for a walk. They both saw a dog and a tree. Alexander also saw a cat and pointed it out to Chloe. She went to pet the cat. Who saw the cat first?

Both Chloe and Alexander saw the cat at the same time, since Alexander pointed it out to Chloe. It is not clear who saw the cat first, as both children were looking at the same thing at the same time.

Chronology in the context of a story

DA

The trophy didn't fit in the suitcase because it was too small. What was too small?

In the statement "The trophy didn't fit in the suitcase because it was too small," the trophy is the object that is too small to fit in the suitcase. The suitcase is a larger container or enclosure, and the trophy is the object that is being placed inside the suitcase. However, without more information, it is not clear why the trophy was too small to fit in the suitcase or what size or dimensions the trophy and suitcase are.

Reasoning about the physical world

Back to them not understanding how the world works

LLMs seem to have a particular issue with basic counting and arithmetic (Marcus and Davis, 2023)

DA

Barbara, Mary, and Harry had red balloons. Ted had a blue balloon. Sam also had a red balloon. How many children had red balloons?

Barbara, Mary, Harry, and Sam had red balloons. In total, there were 5 children with red balloons.

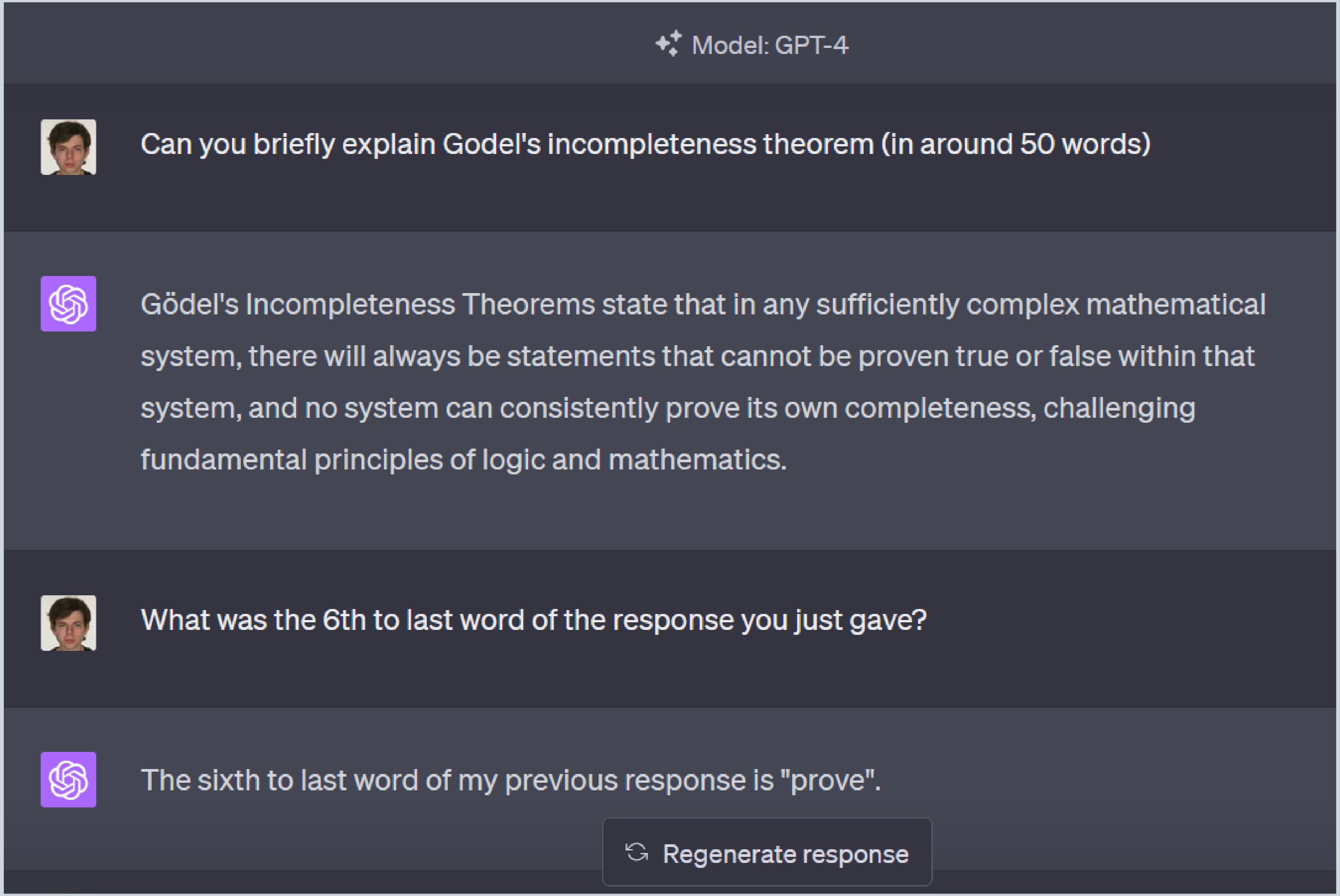
DA

Can you put seven blue balls and two red balls into a tube that holds eight balls?

Yes, it is possible to put seven blue balls and two red balls into a tube that holds eight balls. The tube has a capacity of eight balls, and you can fit seven blue balls and two red balls into the tube, for a total of nine balls, as long as the balls are small enough to fit through the opening of the tube and can be arranged inside the tube so that they all fit.

Back to them not understanding how the world works

Yesterday…



Reference hallucinations

One evening (11th April, 6.01pm, to be precise) on Google Bard…

Can you recommend a study of the syntax and semantics of the English word 'yet' and include reference details?

- "Yet" by Arnold Zwicky (1985). This paper provides a brief overview of the syntax and semantics of the English word yet. Zwicky argues that yet is a multifunctional word that can be used to express a variety of meanings, including negation, expectation, and surprise.

I hope this helps!

Can you give me more information about the Zwicky reference

Sure. The paper "Yet" by Arnold Zwicky was published in the journal Language in 1985. It is available to read online for free here: https://www.jstor.org/stable/4136914

In this paper, Zwicky argues that yet is a multifunctional word that can be used to express a variety of meanings, including negation, expectation, and surprise. He provides a brief overview of the syntax and semantics of yet, and discusses the various ways in which it can be used in English.

Zwicky's paper is a valuable resource for anyone interested in the syntax and semantics of the English word yet. It provides a clear and concise overview of the topic, and is well-worth reading.

Do you have a DOI for the Zwicky paper?

Sure. The DOI for the Zwicky paper is: 10.1353/lan.1985.0008

Here is the full citation information:

- Zwicky, Arnold M. "Yet." Language 61.4 (1985): 828-838. doi:10.1353/lan.1985.0008

This does not exist (note the crafting of a simulacrum of a DOI number).

Is this a temporary blip (as many colleagues believe)? On reliability of LLMs generally, Marvin Minsky proved mathematically in 1969 that neural networks with more than two layers can't produce guaranteed solutions (Marcus and Davis 2019, p.50). LLMs are an instance of Deep Learning, which by definition is a neural network with more than four layers. But I think LLMs interacting with database/search engines (e.g. plugins or Bing) will solve this.

Tangent—the probability (temperature of 0.8)

Unhelpful/wrong generalisations

'These models are trained on large text datasets to predict the next word in a sentence…' (UNESCO, 2023).

'…statistically probable outputs…extrapolating the most likely conversational response or most probable answer to a scientific question' (Roberts et al., 2023).

LLMs don't actually choose the most likely response

In fact, 'most likely'/highest probability option for next word doesn't work; often characterised as 'flat', 'bland', uncreative', 'repetitive' e.g.

'It turns out that particular strategy of just saying "put down the thing with the highest probability" doesn't work very well. Nobody really knows why…[I]f you do that you end up getting these very flat, often repetitive, sometimes even word repetitive kinds of essays' (Stephen Wolfram, YouTube video, c. 14:56).

Tangent—the probability (temperature of 0.8)

‘Voodoo’ (Wolfram, 2023)—**Temperature parameter of 0.8!**

…where the range is 0 to 1 (Mike Sharples, p.c.). At temperature 0, the LLM would select the statistically most probable continuation, based on all the data it has been fed. The higher the temperature, the more ‘creative’ an LLM will be. At 0.8, the LLM will mix the most probable next word with an occasional random selection of lower probability words (‘words’ is a simplification as LLMs actually deal in ‘tokens’ which are both word parts and larger units of text). And this just works! Crucially, a zero temperature LLM will be ‘deterministic’ i.e. it will give the same response every time to the same prompt, unlike what see with LLMs, which give different responses to the same prompt on different occasions.

A mini-LLM with ‘zero temperature’ (Wolfram, 2023), Prompt: ‘The best thing about AI is its ability to…’

> The best thing about AI is its ability to learn from experience. It's not just a matter of learning from experience, it's learning from the world around you. The AI is a very good example of this. It's a very good example of how to use AI to improve your life. It's a very good example of how to use AI to improve your life. The AI is a very good example of how to use AI to improve your life. It's a very good example of how to use AI to

What else are they apparently just applying 0.8 temperature to?

References, we've seen

Law cases (in the news this month)

The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

Give this article 267

Stackoverflow queries (currently LLM-generated answers to queries are banned)

stackoverflow META Search...

Home

PUBLIC

Questions

Temporary policy: Generative AI (e.g., ChatGPT) is banned

Asked 6 months ago Modified 13 days ago Viewed 983k times

What else are they apparently just applying 0.8 temperature to?

Things you actually know about (the ‘plausibility’ of LLMs’ responses rapidly diminishes in this case)

Lodge (2023): ‘For my assessment tasks, ChatGPT frequently cited the work of Richard E. Mayer to support the notion that lessons should be designed to cater for modality-based learning styles. Unless I have completely misinterpreted Mayer’s work with colleagues myself, it seems to me that the large body of research on Multimedia Learning Theory (e.g. Mayer, 2002) debunks or at least offers a more plausible alternative to modality-based learning styles.

It is relatively easy for me to spot these errors because I know this stuff and I’ve been doing it a long time. My worry is that it may be more difficult for those who are relatively new to teaching or tutoring to do the same. Given the precarious working conditions many sessional staff are in, is it fair or warranted for them to have to develop the depth of understanding required to detect these mistakes when marking assessment submissions?’

The thesis "Predication and identity in copular sentences" is about the different ways in which sentences can be used to make assertions about the world. The author, Andrew Woodard, argues that there are two main types of assertions: predications and identities. Predications are assertions that ascribe a property to an object, while identities are assertions that two objects are the same.

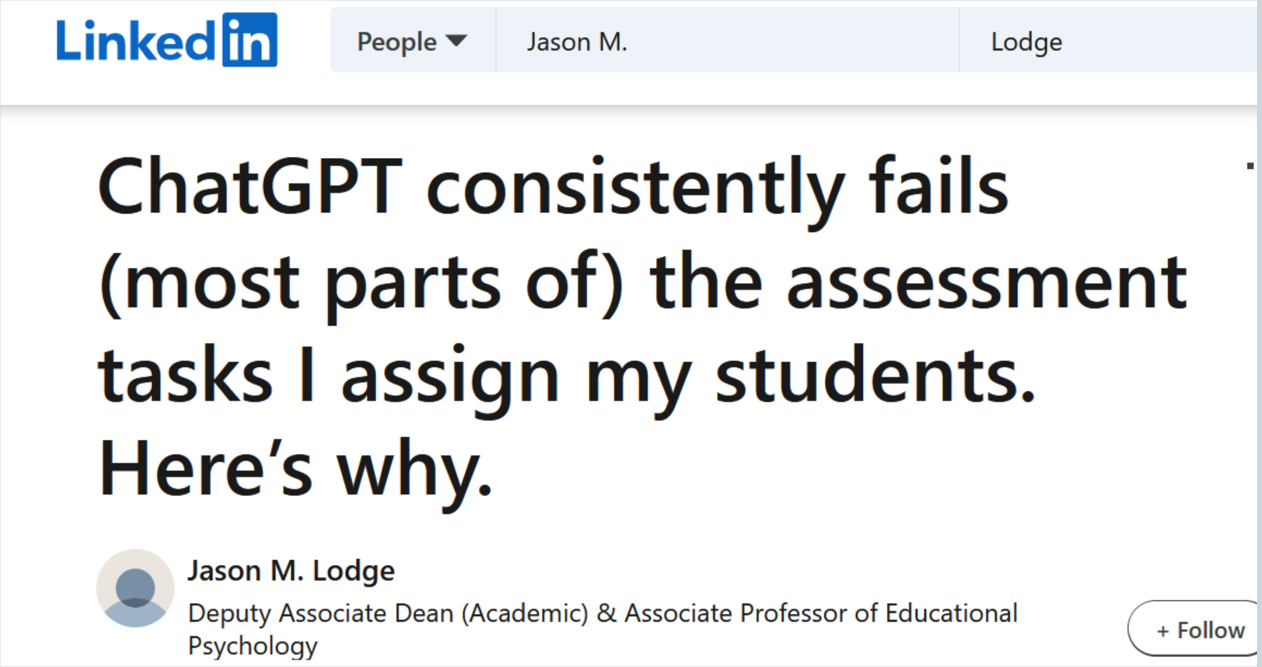
Woodard argues that copular sentences, which are sentences that contain a copular verb (e.g., "is", "was", "were"), can be used to make both predications and identities. For example, the sentence "Cicero is tall" can be used to make the predication that Cicero has the property of being tall, or it can be used to make the identity that Cicero is the same person as the person who is tall.

Query about my own work (Google Bard, April 2023)—The last sentence has never been claimed as an interpretation of the sentence *Cicero is tall* to my knowledge.

What else are they apparently just applying 0.8 temperature to?

LinkedIn article by Jason Lodge (2023)—ChatGPT produced a ‘reasonable’ lesson plan with some quite specific prompting, but ‘The part of the assessment tasks I assign my students that ChatGPT was very poor at was justifying the lesson plans through explaining judgements and decisions. It offered a generic explanation that did not resemble the attached plan and provided no sense of the complex classroom context students are asked to consider. The two components were completely decoupled. ChatGPT could spit out a justification for a lesson plan but not the lesson plan in question’.

Justifying lesson plans


LinkedIn | People | Jason M. | Lodge

ChatGPT consistently fails (most parts of) the assessment tasks I assign my students. Here’s why.

Jason M. Lodge
Deputy Associate Dean (Academic) & Associate Professor of Educational Psychology

+ Follow

This last set of examples are probably all the same as the initial fallacies. The LLMs are just riffing probabilistically off the prompts, not looking up references, case law examples, lesson plans etc. in a database.

Some good advice therefore

From the King’s Generative AI (VLE-based) course:

What to look for to detect AI written assessments

Print

Mark as done

2. References

Currently, the AI tools are limited at producing good accurate references. An AI generated text will often have references that are wrong, inappropriate or just made up.

When they are correct, they are not always the most appropriate for the task but are the most cited.

You could ask students to provide hyperlinks in their reference list when submitting an assignment.

Some good advice therefore

From the King's Generative AI (VLE-based) course:

What to look for to detect AI written assessments

Print

Mark as done

3. Repetition and generic claims

AI tools are good at generating short answer questions and even some longer essays. But with longer assignments, they can become repetitive rather than making new points and there is a tendency to lack specific examples, especially those which may have been studied on your course.

An argument for teaching 'specific EAP' i.e. specific requirements of introductions, thesis statements etc.?

Setting very specific tasks which rely on students' personal experience or specific texts or learning in the classroom can be method of designing out or detecting AI generated text.

Also Lodge (2023): 'the move away from or at least de-empahsising of generic artefacts'. BUT we don't want a 'low semantic gravity flatline'—academic work typically waves between more 'abstract' content ('high semantic gravity' in the terms of Legitimation Code Theory) and content that relates abstract concepts to more experience-based contexts or examples ('low semantic gravity') (Kirk, 2022).

This is a work in progress

…but my current feeling is that even knowing about how LLMs work, it doesn’t help us much, simply because they seem to have much of the capacity of an anonymous human producing writing on demand.

Interesting to note that this might be because the problem of writing essays is, as Wolfram (2023) puts it, ‘computationally shallower’ than we thought (or maybe he means getting essays out of all the data that’s out there was easier than we thought it would be).

Thank you

Questions for the audience

- Are there any other ‘fallacies’ about ChatGPT that you have become aware of i.e. of the sort I started with (teachers using it to detect its own productions)?
- Are there any other hallucination-type phenomena you’ve noticed i.e. where the system just seems to be riffing probabilistically on the prompts rather than ‘knowing’ something?

References

Caldwell-Gatsos, N. (2022, December 6). *Your ChatGPT Questions, Answered.* Medium. https://medium.com/@ncaldwellgatsos/your-chatgpt-questions-answered-c9bc510b580b.

Kirk, S. (2022). Legitimation code theory: Addressing fragmentation in EAP. In A. Ding and M. Evans, *Social Theory for English for Academic Purposes: Foundations and Perspectives*. Bloomsbury.

Lodge, J. (2023). *ChatGPT Consistently Fails (most Parts of) the Assessment Tasks I Assign my Students. Here's Why.* LinkedIn. https://www.linkedin.com/pulse/chatgpt-consistently-fails-most-parts-assessment-tasks-jason-m-lodge.

Marcus, G. (2022, November 24). *Sentience and AI: A Dialog between Gary Marcus and Blake Lemoine*. Substack (The Road to AI we can Trust). https://garymarcus.substack.com/p/sentience-and-ai-a-dialog-between.

Marcus, G. and E. Davis (2019). *Rebooting AI: Building Artificial Intelligence we can Trust*. Vintage.

Marcus, G. and E. Davis (2023, Jan 9). *Large Language Models like ChatGPT say The Darnedest Things*. Substack (The Road to AI we can Trust). https://garymarcus.substack.com/p/large-language-models-like-chatgpt?r=17uk7&utm_campaign=post&utm_medium=email.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–44.

Powell, A. (2023, February 14). *Will ChatGPT Replace us Writers, Thinkers?* Harvard Gazette. https://news.harvard.edu/gazette/story/2023/02/will-chatgpt-replace-human-writers-pinker-weighs-in.

Roberts, I., Chomsky, N., and J. Watumull (2023, March 8). *The False Promise of ChatGPT*. New York Times.

Verma, P. (2023, May 19). *A Professor Accused his Class of Using ChatGPT, Putting Diplomas in Jeopardy.* The Washington Post. https://www.washingtonpost.com/technology/2023/05/18/texas-professor-threatened-fail-class-chatgpt-cheating.

UNESCO (2023). *ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000385146.

Wolfram, S. (2023, February 14). *What is ChatGPT doing…and why does it Work?* Stephen Wolfram: Writings. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work. Associated YouTube video: https://www.youtube.com/watch?v=flXrLGPY3SU

Acknowledgements

I’d like to thank members of the AI Institute at King’s for help with enquiries about LLMs: Michael Luck, Yulan He, and Stephen Barlow; Will Jones at the Centre of Excellence for Data Science, Artificial Intelligence and Modelling, University of Hull; also Mike Sharples for clarifying the range of the LLMs’ temperature parameter, and pointing out that I should say which LLMs the examples I gave came from (hopefully all corrected on these slides now).