Personal Science Guide to AI

Richard Sprague

2024-04-29

Table of contents

1 About

2024-04-28

> Technology moves faster than science.

Although the field of Artificial Intelligence has been around for decades, the recent explosion of Large Language Models (LLMs) is rapidly making professional-grade tools available for personal use.

This site is an up-to-date collection of resources we hope will be useful to personal scientists, those of us who want to use science for personal, rather than professional reasons.

Everything here is currently an early draft, but we update it daily, so if you don't like today's version, come back tomorrow.

If you have questions, comments, or additions please let us know

---

Because at least so far we think a written book is still the best way to consolidate and share high-quality information, we try to split this content into two parts:

1. An up-to-date website
2. A downloadable and printable book for more timeless information.

Both texts are still far from complete, so subscribe for regular updates:

Stay in touch

For weekly updates

2

this is a book version

AI and Creativity

Explaining LLMs

Heidegger and AI

AI Linguistics Perspective

Computing Power and Data Needed for AI

3 How Does it Work?

My explanation of how LLMs do what they do

3.1 How does an LLM-based generative system work?

Imagine you have access to a zillion documents, preferably curated in some way reassures you about their quality and consistency. Wikipedia, for example, or maybe Reddit and other posts that have been sufficiently up-voted. Maybe you also have a corpus of published articles and books from trustworthy sources.

It would be straightforward to tag all words in these documents with labels like "noun", "verb", "proper noun", etc. Of course there would be lots of tricky edge cases, but a generation of spelling and grammar-checkers makes the task doable.

Now instead of organizing the dictionary by parts of speech, imagine your words are tagged *semantically*. A word like "queen", for example, is broken into the labels "female" and "monarch"; change the label "female" to "male" and you have "king". A word like "Starbucks" might include labels like "coffee", as well as "retail store" or even "Fortune 500 business". You can shift the meaning by changing the labels.

Generating a good semantic model like this would itself be a significant undertaking, but people have been working on this for a while, and various good "unsupervised" means have been developed that can do this fairly well.

Auto-completion is a simple form of this. With any sized corpus, you'll know with reasonable probability the likelihood that a particular word will follow another word. Interestingly, you can do this in any human language without even knowing about that language – the probabilities of word order come automatically from the sample sentences you have from that language.

Now go a step above auto-completion and allow for completion at the sentence level, or even the paragraphs or chapters. Given a large enough corpus of quality sentences, you could probably guess with greater-than-chance probability the kinds of sentences and paragraphs that should follow a given set of sentences. Of course it won't be perfect, but already you'd be getting an uncanny level of sophistication.

Pair this autocompletion capability with the work you've done with semantic labeling. And maybe go really big, and do this with even more meta-information you might have about each corpus. A Wikipedia entry, for example, knows that it's about a person or a place. You know which entries link to one another. You know the same about Reddit, and about web pages. With enough training, you could probably get the computer to easily classify a given paragraph into various categories: this piece is fiction, that one is medical, here's one that's from a biography, etc., etc.

Once you have a model of relationships that can identify the type of content, you can go the other direction: given a few snippets of one known form of content (biography, medical, etc.), "auto-complete" with more content of the same kind.

This is an extremely simplified summary of what's happening, but you can imagine how with some effort you could make this fairly sophisticated. In fact, at some level isn't that what we humans are already doing. If your teacher or boss asks you to write a report about something, you are taking everything you've seen previously about the subject and generating more of it, preferably in a pattern that fits what the teacher or boss is expecting.

Some people are very good at this: take what you heard from various other sources and summarize it into a new format.

"List five things wrong with this business plan", you don't necessarily need to *understand* the contents. If you're good enough at re-applying the patterns you've seen from similar projects, you'll instinctively throw out a few tropes that have worked for you in the past. "The plan doesn't say enough about the competition", "the sales projections don't take X and Y into account", "How can you be sure you'll be able to hire the right people". There are thousands, maybe *hundreds of thousands* of books and articles that include these patterns, so you can imagine that with a little tuning a computer could do an exceptional job at this.

3.2 Fine-tuning the output

Simple text-completion will only get you so far[1]. Usable systems need refinement to make them behave more in the way we expect.

Reinforcement learning works by applying a reward or penalty score to the output and then retraining recursively until the model improves to an acceptable level.

Reinforcement learning with human feedback (RLHF) takes this a step further by including humans in the reward formula. The system generates multiple versions of an answer and a human is asked to vote on the best one.

Reinforcement learning with AI feedback (RLAIF) tries to use the AI itself to provide the feedback

see Thomas Woodside and Helen Toner: How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2 for a detailed but readable discussion.

[1] see Karpathy for examples

3.3 Wisdom of the crowds

An LLM is sampling from an unimaginably complex mathematical model of the distribution of human words – essentially a wisdom of crowds effect that distills the collective output of humanity in a statistical way.

3.4 Where do you get the documents

OpenAI gets its documents from more than 200 million documents, 93% of which are in English, that are selected to be representative of a broad space of human knowledge.

Of course it starts with Wikipedia: almost 6 million articles.

One set of words comes from Common Crawl: a large, public-domain dataset of millions of web pages.

Another is a proprietary corpus called *WebText2* of more than 8 million documents made by scraping particularly high-quality web documents, such as those that are highly-upranked on Reddit.

Two proprietary datasets, known as *Books1* and *Books2* contain tens of thousands of published books. These datasets include classic literature, such as works by Shakespeare, Jane Austen, and Charles Dickens, as well as modern works of fiction and non-fiction, such as the Harry Potter series, The Da Vinci Code, and The Hunger Games.[2] There are also many other books on a variety of topics, including science, history, politics, and philosophy.

> Also high on the list: b-ok.org No. 190, a notorious market for pirated e-books that has since been seized by the U.S. Justice Department. At least 27 other sites identified by the U.S. government as markets for piracy and counterfeits were present in the data set.

Washington Post has an interactive graphic that digs into more detail. (Also discussed on HN)

Yes, they crawl me:

[2] see Apr 2023 Chang et al. (2023)

Figure 3.1: blog.richardsprague.com tokens on Google's C4 dataset

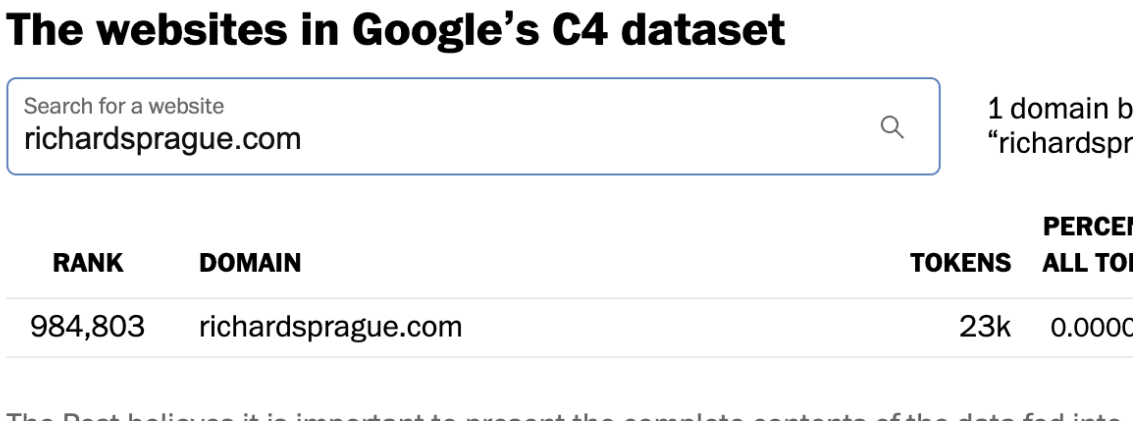

Figure 3.2: richardsprague.com tokens on Google's C4 dataset

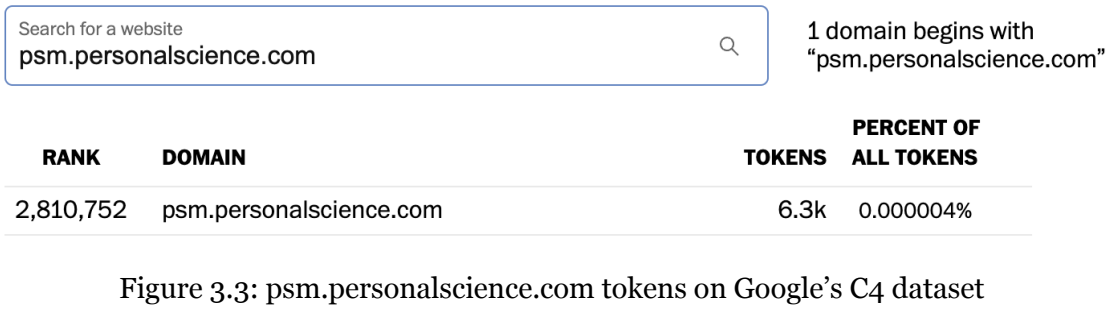

Figure 3.3: psm.personalscience.com tokens on Google's C4 dataset

from NYTimes

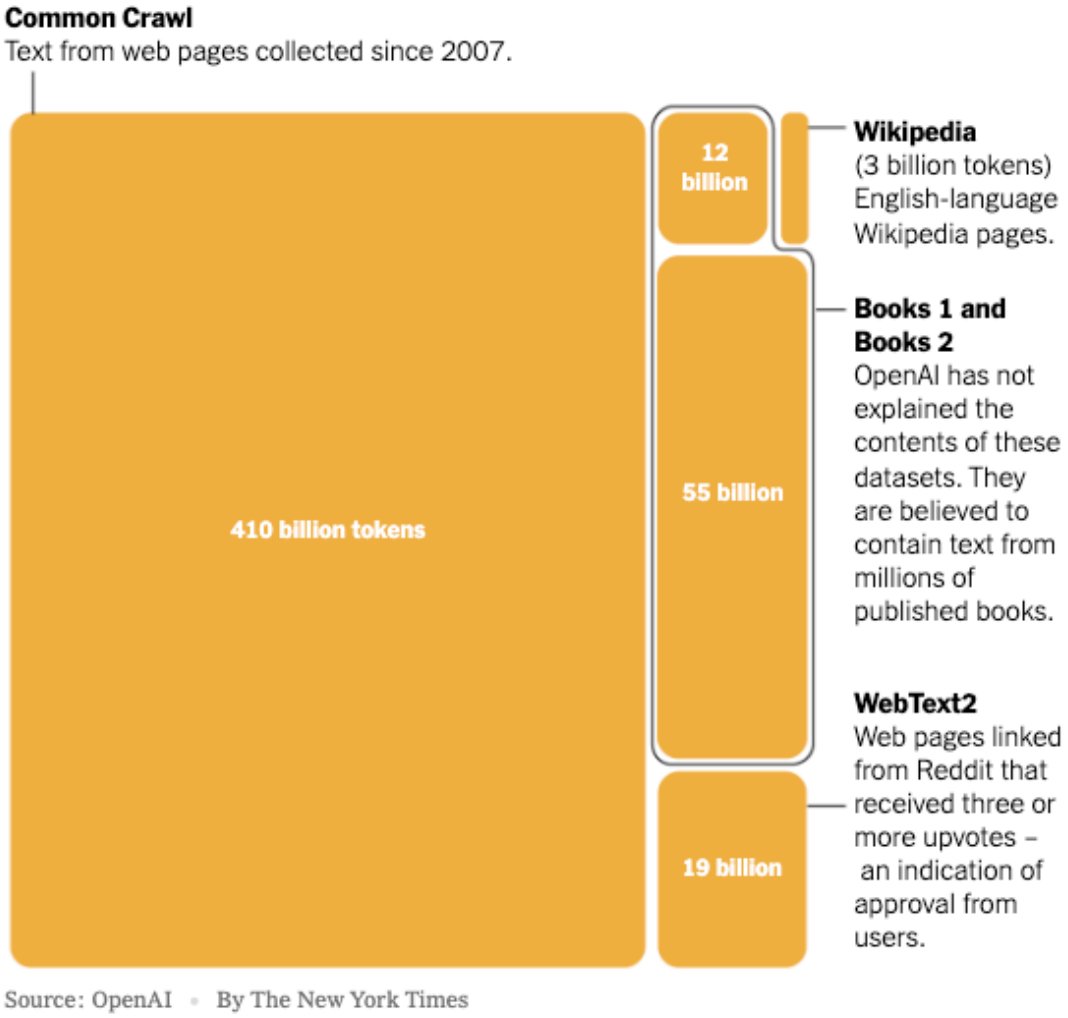

Figure 3.4: GPT-3 Data Sources

3.5 More Data

There is another dataset, *Books3* originally intended to be an open source collection of books.

See Battle over books about books3 (Wired) and another from The Atlantic Revealed: The Authors Whose Pirated Books are Powering Generative AI

> a direct download of "the Pile," a massive cache of training text created by EleutherAI that contains the Books3 dataset, plus material from a variety of

> other sources: YouTube-video subtitles, documents and transcriptions from the European Parliament, English Wikipedia, emails sent and received by Enron Corporation employees before its 2001 collapse, and a lot more

The Pile is defined and described in this 2020 paper at arXiv.

See more details on HN, including direct links to the latest downloads.

You can download your own store of over 200 million OCR-scanned pages at HuggingFace US-PD-books

Peter Schoppert writes much more about the copyright issues involved in The books used to train LLMs

3.6 Organizing the data

At the core of a transformer model is the idea that many of the intellectual tasks we humans do involves taking one sequence of tokens – words, numbers, programming instructions, etc. – and converting them into another sequence. Translation from one language to another is the classic case, but the insight at the heart of ChatGPT is that question-answering is another example. My question is a sequence of words and symbols like punctuation or numbers. If you append my question to, say, all the words in that huge OpenAI dataset, then you can "answer" my question by rearranging it along with some of the words in the dataset.

The technique of rearranging one sequence into another is called Seq2Seq*.

3.7 References and More Details

Carlini et al. (2024) > You can steal portions of closed language models such as the embeddings layer just by using their public APIs. This can be done for a modest budget of less than $2,000.

Dodge et al. (2021)

An in-depth way to study what data was used to train language models. The results of which suggest that many closed-source models likely didn't train on popular benchmarks:

Oren et al. (2023)

> we audit five popular publicly accessible language models for test set contamination and find little evidence for pervasive contamination.

4 AI and Creativity

4.1 AI and Creativity

How creative is AI?
Ethan Mollick summarizes the research showing that AI is more creative than humans (though not as good as the most creative people). He recommends prompts that constrain the AI (e.g. limit the number of answers, require it to focus on a particular problem-solving style).

Using ChatGPT for Historical Simulations

The fact that LLMs hallucinate can be used to advantage in many contexts, such as education, where often people learn better by simulating a variety of situations.

Benjamin Breen offers The Case for LLMs as Hallucination Engines with many examples including links to sessions you can do with ChatGPT.

For example, The Fall of the Ming Dynasty is a history simulation for ChatGPT, where you are given a realistic description of a specific setting and asked to react. Your response leads to the next setting, and so on.

The Future of Poetry

> 38 AI experts and 39 English experts were asked to rate and guess whether poems were written by an AI or a human. The human came in 1st place, while Bard, ChatGPT-4, and Claude came in 2nd, 3rd, and 4th places respectively, both in writing quality and their ability to fool respondents into believing their poems were authored by a human. English experts were far better at discerning which poems were written by AI, which points to a need for them to play a greater role in helping shape future versions of AI technology.

In February 2023, The Atlantic and The Washington Post examined AI poetry, concluding that AI poems were clichéd, predictable, and full of awkward rhymes.

5 Philosophy

see AI and Philosophy and Religion

"AI is applied philosophy": so said my Stanford professors back when they were creating what came to be the popular "Symbolic Systems" major. Trying to understand the implications of this new breed of computers requires pulling across disciplines from neuroscience to math to psychology and more – perfect for the back-to-fundamentals questions that philosophers have asked for thousands of years.

see also Can Machines Think

5.1 Stephen Wolfram

In What is ChatGPT Doing and How Does it Work?:

> And in the end there's just a fundamental tension between learnability and computational irreducibility. Learning involves in effect compressing data by leveraging regularities. But computational irreducibility implies that ultimately there's a limit to what regularities there may be.

"language is at a fundamental level somehow simpler than it seems"

> but my strong suspicion is that the success of ChatGPT implicitly reveals an important "scientific" fact: that there's actually a lot more structure and simplicity to meaningful human language than we ever knew—and that in the end there may be even fairly simple rules that describe how such language can be put together.

He uses an example of a "balanced parenthesis" language and show how training on 10 million examples or so is enough to reach the competence of humans, who can 'eyeball' a list of parentheses to tell if they're balanced. This shows that human languages have a syntactic constraint that limits the types of token strings they'll accept.

But importantly, note that ChatGPT is also heading toward a semantically *meaningful* constraint as well: it's very good at generating only strings that make sense, i.e. that somehow cohere to whatever humans might accept. Syllogistic logic is one such constraint: sentences that "make sense" must be internally consistent.

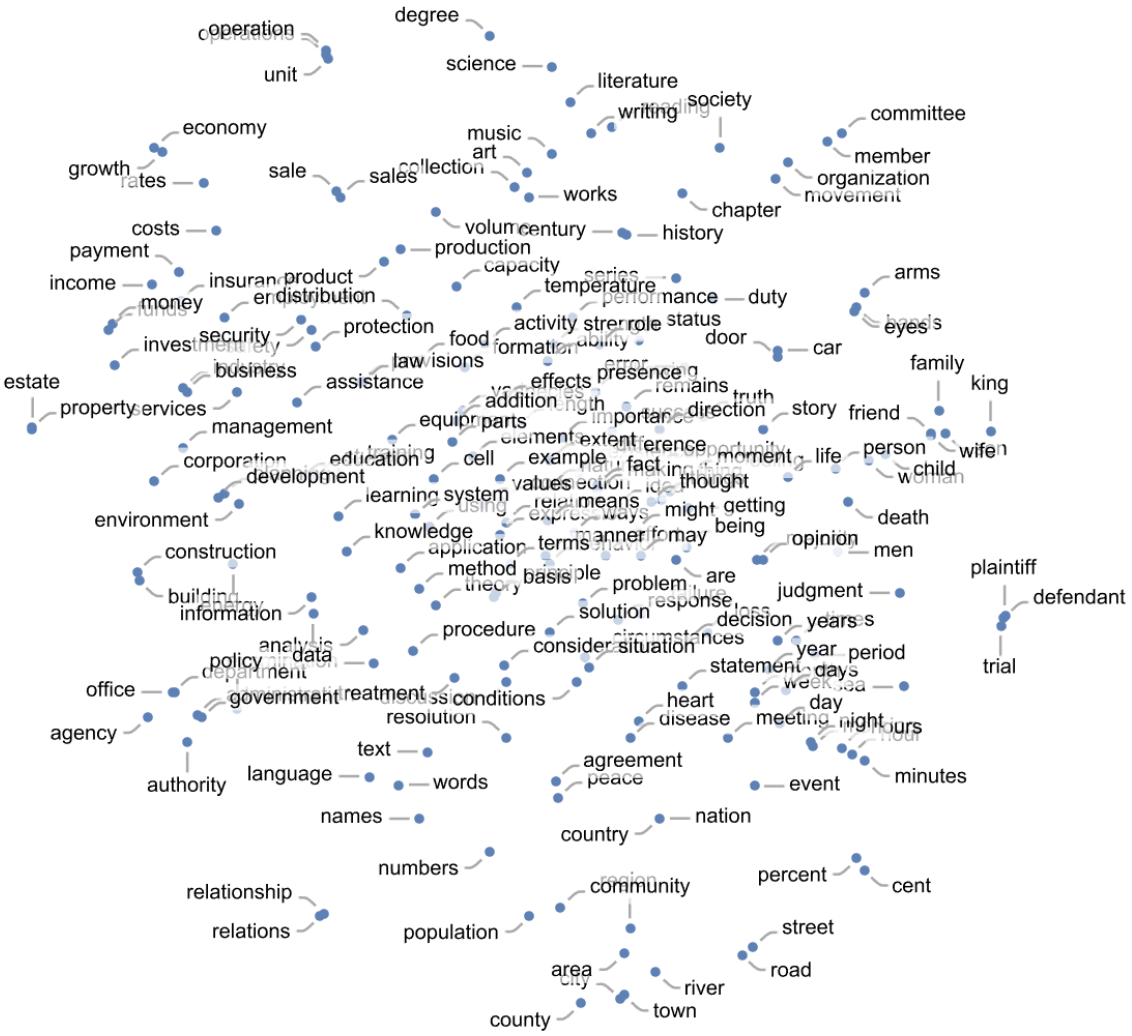

Figure 5.1: Wolfram's example of "meaningfulness space"

Although he doesn't conclude with a generalized theory of meaningfulness, he points to computational languages as proof that it is possible to establish precise constraints – i.e. what can be computed – on what might otherwise be an intractable problem.

July 2023 see his Generative AI Space and the Mental Imagery of Alien Minds

---

François Chollet describes How I think about LLM Prompt Engineering as a search, not just through *word vectors* but also through *vector programs* that correspond to more than simple relationships between words. `Word2Vec` is a 2013-era technology that allows for "arithmetic on words", with operations like `male_to_female(king) -> queen`. Transformers add the concept of *attention* to let these arithmetic operations scale to handle entire sections of text that point to the transformations themselves. Now the search space includes "vector programs" like `write_in_the_style_of_shakespeare(my_input)` and everything in between.

5.2 LLMs and Language

On Stochastic Libraries: Large Language Models as library-like Knowledge-Synthesizers. by René Walter

> These algorithmic machines that are currently emerging are not intelligent, but they provide access to synthetic recombinations of human knowledge, reminiscent of a new type of library whose combined content is navigated by text input. This new form of library doesn't provide access to individual human works, but rather continuations, amalgams, remixes, modulations, and interpolations between nodes in a network of existing knowledge, "in the style of Banksy, trending on artstation."

Jaron Lanier thinks[1], instead of the misleading term "artificial intelligence", we should talk about "an innovative form of social collaboration".

He adds that the future should include a "Mediator of Individual Data (MID)" or data trust, where people pool their content somehow so they can receive some reward when it's reused elsewhere. This "data dignity" is supposed to let the little guy have more say in how his data is used, but I think Lanier is ignoring the *reason* the little guy has no say right now: a single individual contributes very little, in the grand scheme of things. It's only by aggregating lots of tiny contributions that LLMs or other technologies become valuable.

Talking About Large Language Models Shanahan (ICL): Jan 2023

> a great many tasks that demand intelligence in humans can be reduced to next token prediction with a sufficiently performant model.

> **Note**
>
> A bare-bones LLM doesn't "really" know anything because all it does, at a fundamental level, is sequence prediction.

[1] *The New Yorker* April 2023

Language is an aspect of human *collective* behavior

LLMs don't have what Daniel Dennett calls *intentional stance*, i.e. an assumption that an interlocutor has beliefs, desires, or intentions.

Be careful to distinguish between knowing that certain words follow each other and *knowing* a concept.

What if we had a model that used real-time visual feedback from the world in order to update itself. Could such a thing be meaningful described as "knowing"?

The author's answer is "not really", because the visual feedback – while important – is just one narrow part of the world. The system gets *correlational* feedback about that image, sure, but not about how that image fits into the wider world – a *causal* association that humans take for granted.

5.3 Chomsky on AI

2023-01-26 5:32 PM

Also see Chomsky and Moro on the Limits of Our Comprehension, an excerpt from Chomsky and Moro's 2022 book "The Secrets of Words."

"It is important to learn to be surprised by simple facts"

Like the way it is impossible to teach a rat to follow a prime number maze, there may be concepts beyond the ability of human minds to comprehend.

Chomsky wrote a New York Times Op Ed (Mar 2023) Noam Chomsky: The False Promise of ChatGPT in which he gives many classic examples showing how the LLM doesn't really understand anything.

> Note, for all the seemingly sophisticated thought and language, the moral indifference born of unintelligence. Here, ChatGPT exhibits something like the banality of evil: plagiarism and apathy and obviation. It summarizes the standard arguments in the literature by a kind of super-autocomplete, refuses to take a stand on anything, pleads not merely ignorance but lack of intelligence and ultimately offers a "just following orders" defense, shifting responsibility to its creators.

5.4 Gary Marcus on Chomsky

In a 2012 New Yorker piece, Marcus gives some anecdotes about Chomsky:

> This conception of 'renouncing beliefs' is very odd, as if we're in some kind of religious cult. I 'renounce beliefs' practically every time I think about the topics or find out what someone else is thinking."

In Marcus' Substack, explains that Chomsky doesn't like GPT-3 because it's not real science. It doesn't explain *why* it does what it does. It has little explanatory value.

One clue is how GPT-3 is no better at human languages than it is at computer languages. It's all just pattern recognition.

> even in an immense neural network, with hundreds of billions of parameters, performance on simple 3-digit math problems topped out at 80%.

Ask it why it's good to eat socks after meditation.

> the latest and greatest, InstructGPT, was recently asked to explain why it is good to eat socks after meditation and blithely invoked fictitious authorities, alleging that "Some experts believe that the act of eating a sock helps the brain to come out of its altered state as a result of meditation."

also Freddie DeBoer: AI, Ozymandias:

> The human mind is not "a lumbering statistical engine for pattern matching, gorging on hundreds of terabytes of data and extrapolating the most likely conversational response or most probable answer to a scientific question," as Chomsky, Ian Roberts, and Jeffrey Watumull argued earlier this year. The mind is rule-bound, and those rules are present before we are old enough to have assembled a great amount of data. Indeed, this observation, "the poverty of the stimulus" – that the information a young child has been exposed to cannot explain that child's cognitive capabilities – is one of the foundational tenets of modern linguistics.

5.5 AI is Becoming a Natural Science

Former AAAI President Subbarao Kambhampati thinks AI is becoming an "ersatz natural science" more concerned with explaining an empirical phenomenon than with describing the underlying rules.

> A dear colleague of mine used to preen that he rates papers—including his own—by the ratio of theorems to definitions

Remember that Herb Simon used to refer to "sciences of the artificial".

Also see the work of Stanford's Pat Langley, who spent much of his career trying to build machines that can do science.

5.6 Counterargument

Sam Hammond articulates the counter-argument in We're all Wittgensteinians now: The philosophical winners from LLMs (evernote)

> My view is thus the exact opposite of Noam Chomsky's, who argues that the success of Large Language Models is of limited scientific or philosophical import, since such models ultimately reduce to giant inscrutable matrices. On the contrary, the discovery that giant inscrutable matrices can, under the right circumstances, do many things that otherwise require a biological brain is itself a striking empirical datum — one Chomsky chooses to simply dismiss *a priori*.

5.7 Other

Do models use English as their internal language? Paper says it is more that they think in concepts, but that those concepts are biased towards English, so yes they think in English but only in a semantic sense.

AI can be an accessory in the death of traditional languages, or a tool for preserving them.

References

Carlini, Nicholas, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, et al. 2024. "Stealing Part of a Production Language Model." arXiv. http://arxiv.org/abs/2403.06634.

Chang, Kent K., Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4." https://doi.org/10.48550/ARXIV.2305.00118.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." https://doi.org/10.48550/ARXIV.2104.08758.

Oren, Yonatan, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2023. "Proving Test Set Contamination in Black Box Language Models." arXiv. http://arxiv.org/abs/2310.17623.