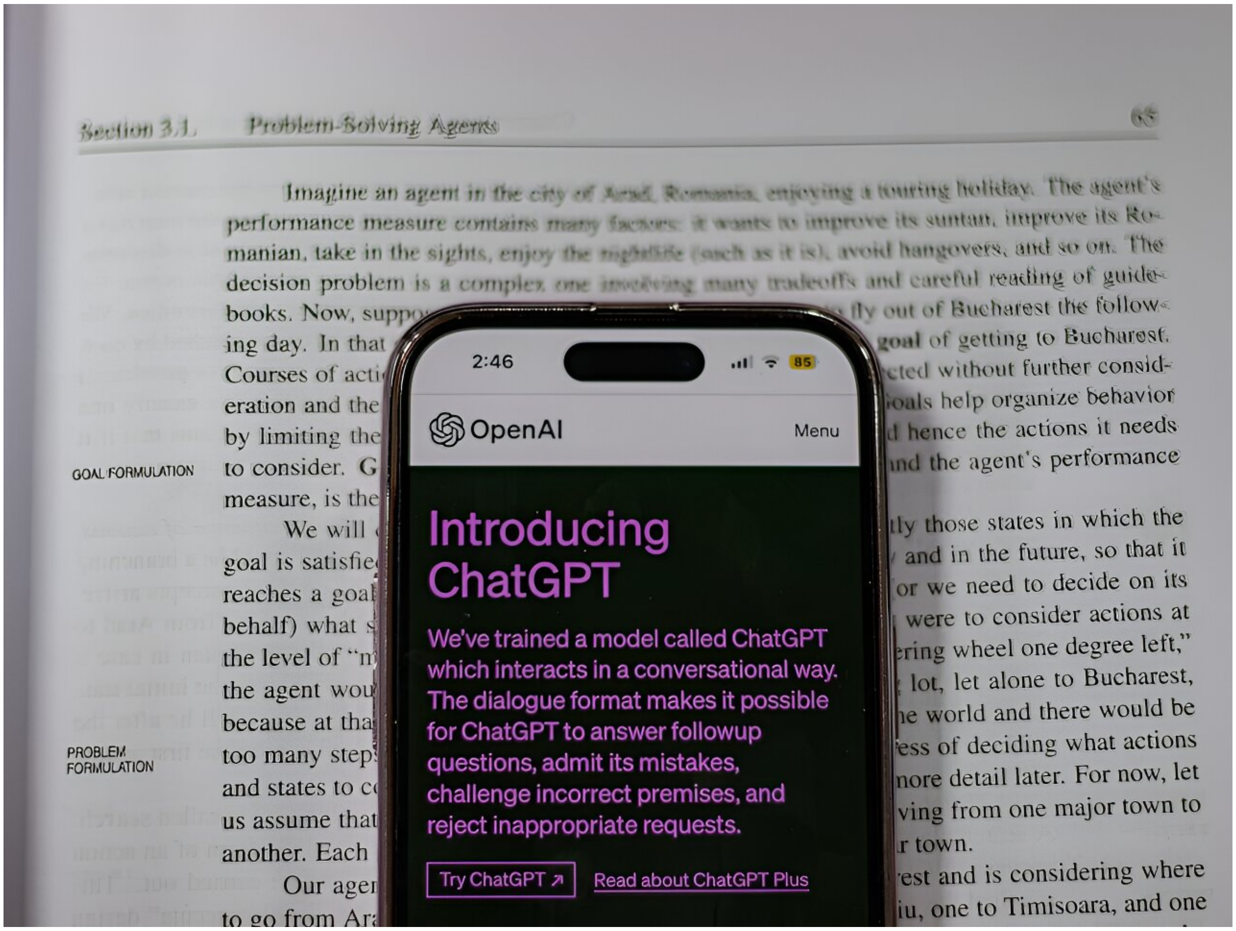
TechXplore

Q&A: Language models—a guide for the perplexed

January 10 2024, by Stefan Milne


Section 3.1. Problem-Solving Agents
65
GOAL FORMULATION
PROBLEM FORMULATION
2:46
OpenAI
Menu
Introducing ChatGPT
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.
Try ChatGPT ↗
Read about ChatGPT Plus

A team University of Washington researchers have published a guide explaining language models, the technology that underlies chatbots. Credit: Shantanu Kumar/Unsplash

Language models have, somewhat surreptitiously, dominated news for the last year. Often called "artificial intelligence," these systems underlie chatbots like ChatGPT and Google Bard.

But a team of researchers at the University of Washington noticed that, even amid a year of AI commotion, many people struggle to find accurate, comprehensible information on what language models are and how they work. News articles frequently focus on the latest advances or corporate controversies, while research papers are too technical and granular for the public.

So recently, the team published "Language Models: A Guide for the Perplexed," a paper explaining language models in lay terms. It is available on the *arXiv* preprint server.

For answers to some common questions, UW News spoke with lead author Sofia Serrano, a UW doctoral student in the Paul G. Allen School of Computer Science & Engineering; co-author Zander Brumbaugh, a masters student in the Allen School; and senior author Noah A. Smith, a professor in the Allen School.

Briefly, what are language models and how do they work?

Serrano: A language model is essentially a next-word predictor. It looks at a lot of text and notices which words tend to follow after which sequences of other words. Typically, when we're talking about a language model, we're now talking about a large machine learning model, which contains a lot of different numbers called parameters. Those numbers are tweaked with each new bit of textual data that the model is trained on.

The result is a giant mathematical function that overall is pretty good at predicting which words come next, given the words that have been supplied in a prompt, or that the model has produced so far. It turns out that these large models also pick up things about the structure of language and things that fall under the umbrella of common sense or world knowledge.

In the paper you bring up this idea of the 'black box,' which refers to the difficulty in knowing what's going on inside this giant function. What, specifically, do researchers still not understand?

Smith: We understand the mechanical level very well—the equations that are being calculated when you push inputs and make a prediction. We also have some understanding at the level of behavior, because people are doing all kinds of scientific studies on language models, as if they were lab subjects.

In my view, the level we have almost no understanding of is the mechanisms above the number crunching that are kind of in the middle. Are there abstractions that are being captured by the functions? Is there a way to slice through those intermediate calculations and say, "Oh, it understands concepts, or it understands syntax"?

It's not like looking under the hood of your car. Somebody who understands cars can explain to you what each piece does and why it's there. But the tools we have for inspecting what's going on inside a language model's predictions are not great. These days they have anywhere from a billion to maybe even a trillion parameters. That's more numbers than anybody can look at. Even in smaller models, the numbers don't have any individual meaning. They work together to take that previous sequence of words and turn it into a prediction about the next word.

Why do you distinguish between AI and language models?

Serrano: "AI" is an umbrella term that can refer to a lot of different research communities that revolve around making computers "learn" in some way. But it can also refer to systems or models that are developed using these "learning" techniques. When we say "language model," we're being more specific about a particular concept that falls under the umbrella of AI.

Smith: The term "AI" brings with it a lot of preconceived ideas. I think that's part of why it's used in marketing so much. The term "language model" has a precise technical definition. We can be clear about exactly what a language model is and is not, and it isn't going to bring up all these preconceptions and feelings.

Serrano: Even within natural language processing research communities, people talk about language models "thinking" or "reasoning." In some respects that language makes sense as shorthand. But when we use the term "thinking," we mostly know how that works for humans. Yet when we apply that terminology to language models, it can create this perception that a similar process is happening.

Again, a language model is a bunch of numbers in a learned mathematical function. It's fair game to say that those numbers are capable of recovering or surfacing information that the model has seen before, or finding connections between input text. But often there's a tendency to go further and make assumptions about any kind of reasoning the models might possess. We haven't really seen this level of fluency decoupled from other aspects of what we consider intelligence.

So it's really easy for us to mistake fluency for all of the other things that we typically roll into the term "intelligence."

Could you give an example of how that fluency translates to things that would be perceived as intelligent?

Brumbaugh: I think determining what a display of intelligence is can be quite difficult. For example, if someone asked a model, "I'm struggling and feeling down—what should I do?" The model may offer seemingly reasoned advice. Someone with limited experience with language models might perceive that as intelligence, instead of next-word prediction.

Smith: If you tell a model, "I'm having a bad day," and its response sounds like a therapist, it has likely read a bunch of articles online that coach people on empathy, so it can be very fluent when it's latching on to the right context. But if it starts feeding on your sadness and telling you you're awful, it's probably latching on to some other source of text. It can reproduce the various qualities of human intelligence and behavior that we see online. So if a model behaves in a way that seems intelligent, you should first ask, "What did it see in the training data that looks like this conversation?"

What makes compiling a good data set to train a language model difficult in some instances?

Brumbaugh: Today's models roughly comprise the entire public internet. It takes enormous amounts of resources to be able to gather that data. In language modeling, essentially, what you put in is what you're going to get out. So people are researching how to best collect data, filter it and make sure that you're not putting in something that's toxic or harmful or just at its lowest quality. Those all present separate challenges.

Why is it vital to have testing data that's not in the original training data set?

Smith: I call this the cardinal rule of machine learning. When you're evaluating a model, you want to make sure that you're measuring how well it does on something it hasn't seen before. In the paper, we compare this to a student who somehow gets a copy of the final exam answer key. It doesn't matter whether they looked at it. Their exam is just not useful in judging whether they learned anything.

It's the same with language models. If the test examples were in the training data, then it could have just memorized what it saw. There's a large contingent of researchers who see these models as doing a lot of memorization—maybe not perfect memorization, but fuzzy memorization. Sometimes the word "contamination" gets used. If the training data was contaminated with the test, it doesn't mean the language model is stupid or smart or anything. It just means we can't conclude anything.

What's it important for the public to understand about language models right now?

Brumbaugh: We need to keep separating language models from notions of intelligence. These models are imperfect. They can sound very fluent, but they're prone to hallucinations—which is when they generate erroneous or fictional information. I know people who are using language models for something relatively important, such as looking up information. But they give a fuzzy representation of what they've learned. They're not databases or Google search.

Smith: If you look at great technological achievements—the airplane or the internet—most resulted from having a clear goal. We wanted to move people through the air, or send information between computers. But just a few years ago, language models were largely research artifacts. A few were being used in some systems, such as Google Translate. But I don't think researchers had a clear sense of solving a problem by creating a product. I think we were more saying, "Let's see what happens if we scale this up." Then, serendipitously, this fluency yielded these other results.

But the research wasn't done with a target in mind, and even now nobody quite knows what that target is. And that's kind of exciting because some of us would like to see these models made more open because we think there is a lot of potential. But big tech companies have no reason to make a tool that works really well for Sofia or me or you. So the models have to be democratized.

What are some basic steps toward that democratization?

Smith: Some organizations are building language models that are open, where the parameters, code and data are shared. I work part-time for one of those organizations, the Allen Institute for Artificial Intelligence, but there are others. Meta has put out models, without the data, but that's still better than nothing. A company called EleutherAI puts out open models. These models are still often quite expensive to run. So I think we need more investment in research that makes them more efficient, that lets us take a big model and make it cheap enough to run on a laptop.

More information: Sofia Serrano et al, Language Models: A Guide for the Perplexed, *arXiv* (2023). DOI: 10.48550/arxiv.2311.17301

Provided by University of Washington

Citation: Q&A: Language models—a guide for the perplexed (2024, January 10) retrieved 14 May 2024 from https://techxplore.com/news/2024-01-qa-language-perplexed.html