Memo on Generative Language Models

Scott Robert Patterson
2 March 2023

Summary

This memo details recent developments in generative language models as they relate to pedagogy and research. The report provides background context for these developments, notes how these developments may impact assessment, and details several examples of how this technology can be used for teaching and research.

Part 1: Context

What is ChatGPT?

The release of the ChatGPT research test preview in November 2022 marked - brought a new level of public attention to the AI world. Developed by OpenAI, which has strong connections with Microsoft, it is a generative pre-trained language model, which means that it is trained to predict a sequence of words based on a prior sequence of words.

ChatGPT is a brand name with several components. First, “GPT” stands for *Generative Pre-trained Model*. The model is “generative” because it can generate text based on an input sequence - or “prompt.” “Pre-trained” describes the form of the model, meaning that OpenAI compiled a massive collection of text data and used machine learning to learn the mathematical relationships between words. Although GPT is a brand name, a more generic name for this type of model is a “foundational” model. The “Chat” in ChatGPT refers to a post-training modification to a language model that is “fine-tuned” to answer questions in a human-like way.

While pre-trained language models are static, it is possible to build additional layers and functionalities on top of them. ChatGPT is an example of such modifications. The model is fine-tuned to answer questions and respond to prompts more like a human.

There are two main reasons why ChatGPT has attracted so much hype. Firstly, the "chatting" feature makes it easier to prompt the model, which generates text based on the prompt. Secondly, it is free to use.

foundational Models

The current era of machine learning began in 2017, with the publication of the "transformer" model architecture. This architecture improved upon the previous state-of-the-art in two key ways. Firstly, it increased the attention span of models by allowing them to remember longer sequences of words when making predictions. In contrast, older models were limited to short sequences of words, leading to less coherent writing. Secondly, the transformer architecture allowed for more efficient processing of training data by distributing the job over multiple processing units. This allowed for an explosion in the size of training datasets and led to research focusing on discovering new functionalities that emerge from scale and finding ways to connect language generating models to other technologies.

At the moment, the capabilities frontier of foundational models is set by their scale. In brief, increases to the quantity of training data appear to produce new qualities of capabilities. In a sense, much debate about the internal limits of foundational models are debates about the functional limits of scale. If there is a limit to how much scaling up can yield new functionalities, most experts agree that it has not yet been reached.

That said, the question of what language models can do is not a factor of their internal limitations alone. In fact, there are considerable research and commercial efforts that aim to integrate foundational models with other technologies. Some examples include general and scholarly search engines, as well as knowledge graphs. Overall, it is important to note that new capabilities of language models may emerge from both internal technological improvements and from combination with other technologies.

Who are the key actors?

The key division among companies working in the field is between the owners and renters of computing infrastructure. The reasons why are because of 1) capital costs and 2) because these are 'general' models.

1. Capital Costs: Training a transformer model is very expensive: assembling the hardware, expertise, and data to build such a model is extremely expensive. Training a single iteration of GPT-3 actually costs millions of dollars. Additionally, the hardware used to generate these models is currently at the center of the US-Chinese trade war and was also seriously disrupted by pandemic supply chain issues. This means that a few companies have a first-mover advantage with path dependent lock-in.
2. General-purpose language models: the scale of these models also means that they are suitable for a wide variety of tasks. While the pre-trained element of a model is static after training, these models can be built upon to perform more specific tasks. If you recall ChatGPT, this is a general purpose model that has been fine-tuned for the task of question answering. This means that once a company builds a model, they can offer access to the model as a service.

These factors have consolidated market power among a few key actors (most notably Open AI - linked to Microsoft - and Google).

Many smaller startup companies build niche applications by accessing the larger foundational models. At the moment, there are hundreds of companies that offer AI writing assistant tools. Many other companies aim to link foundational models, like GPT-3, to other software services that are offered by API. For example, 'perplexity.ai' is a foundational model linked to a general purpose search engine. 'Elicit' and 'Consensus' offers a foundational model that is linked to a scholarly search engine. In practice, this looks like doing a Google search, but instead of a list of answers, natural language is returned. Like all foundational models, factuality can be suspect. However, by linking to a search engine, these models do return sources so that the user can verify. On March 1, OpenAI announced the launch of the ChatGPT API. This means that third-party developers can now integrate ChatGPT into their own applications. The majority of these smaller companies are linked to OpenAI's API. At the moment, it is unclear whether Google will opt for a similar strategy.

OpenAI released ChatGPT for free as a research preview, which was an expensive marketing decision. In contrast, Google also has a foundational language model and has been actively developing related technology for a while. However, they have chosen to keep their models very constrained and only available for closed beta-testing due to ethical concerns. They claim that their cautious approach is necessary, but they are now revisiting that strategy in light of ChatGPT's success. For example, Google announced "Bard" on February 6, 2023 - which is billed as a rival to ChatGPT. In February, Microsoft also announced the integration of similar features to its Bing search engine. This is currently available in a closed preview. One distinguishing feature of the Microsoft offer is that its bot is connected to the internet, unlike ChatGPT. In addition to these examples, several other companies have announced similar models, but the number is small.

Large tech companies have invested heavily into language models at a time of high interest rates and of widespread layoffs in the tech sector. Both Google and Microsoft have expressed intent to integrate features of generative language models into their core business, with Microsoft making very public efforts to do so fast. This could look like an integration of these into Microsoft Office or Google Drive in a few years. It is also possible that the broad availability of these tools now may reflect an environment where large tech companies are incentivizing start-ups to discover new uses for the technology that may eventually be integrated and mainstreamed by the big players.

Part 2: Impact on Assessment

Access to language models

Students currently have access to a variety of AI writing assistant tools, most of which are offered through a ‘freemium’ subscription offer. Most of these models are linked to the OpenAI API, but there may soon be other options. Maintaining these models is expensive, so it is possible that access may become more expensive down the road. That said, in the short term, most students will be able to access some form of AI assistance and savvy students will be able to access more niche tools. For the foreseeable future, it should be expected that students will have access to these tools.

The availability and cost of language models may change in the future, particularly if venture capital funding dries up or if enterprises establish more reliable customer bases. Additionally, it is likely that language models will become more fine-tuned to specific tasks or fields over time. While this could lead to more efficient and effective models for various disciplines, it could also create potential issues for students.

If language models become more expensive, it could lead to an unfair advantage for students who can afford them. This could potentially widen existing socioeconomic disparities in education. On the other hand, if developers create highly effective, discipline-specific language models, it could affect certain disciplines more than others. It is possible that there could be an effective, discipline-specific language model available at a high cost, which could also contribute to disparities in education. Overall, the changing landscape of language models has the potential to impact students in various ways, both positive and negative.

Uses & Limitations

Text generation is the most prominent use case for these models. Text generation is performed through "prompting" - where the user enters a text sequence and the model generates text based on the input. Prompting can be done in one go (zero-shot prompting) or can be performed sequentially/imitatively (few-shot prompting). Language models can be used for many other tasks, such as building question-answering applications, classifying texts based on their features, or annotating (for example, finding the research question of an article). While these uses may eventually have impacts on assessment, they are less immediate.

Two of the most important limitations of language models are their 1) factual inaccuracy and their 2) limited attention span.

1) Factual Inaccuracy: language models tend to make lots of factual errors and confidently fabricate answers. In the community of practice, this is called "hallucination." Language models are very bad with citations - they are often non-existent. The potential work around for this comes when language models are linked to search engines and are thereby able to return source material. Still, this problem is unlikely to be solved in the short term.
2) Language models are limited in their "attention span" - that is, how many words of context they can keep in memory. With GPT-3, the limit is somewhere between 800-1500 words. This word limitation also includes the text used for prompting. Therefore, I can not take the text of a book (or even a journal article) and generate a summary with the push of a button.

In practice, these factors mean that a student is unlikely to be able to generate anything useful without actually engaging with source material. A zero-shot essay would likely be riddled with factual inaccuracies and would meander off topic. To successfully use a language model, a student would likely need to provide their own notes, break the problem down into smaller prompts, generate small bits of text, verify and fact check, and then knit them together in a sequence.

At the moment, there are limits to the maximum length of coherent text generation. With GPT-3, the limit is somewhere between 1000-1500 words. In practice, this means that once the limit is hit, the model will no longer remember what was written at the beginning. The upshot is that a student is not likely to be able generate a passable essay with the push of a button. A zero-shot essay would likely be riddled factual with errors and off topic. A successful use of a language model would likely involve breaking the problem down into smaller questions, generating text for those smaller questions, verifying/fact-checking, and editing.

There are a few potential cases where zero-shot or few-shot prompting could generate a passable essay. For example, if an assignment asks for a question that is widely asked, it is possible that similar answers may appear in the training data of a language model. However, this is only likely if the same question has been recycled broadly for many years.

AI plagiarism detection

There are new programs available that claim to be able to detect whether a text was generated by a language model. However, these programs are still very new and it is unclear how they work exactly. One thing that is likely is that any detection will be probabilistic, meaning that the output of the detection will be a likelihood that something was generated by a language model. It is uncertain how this would be operationalized in a detection app, but it is unlikely that anyone can say for certain whether something was artificially generated or not.

While there are efforts being made to create detection software, there are far more resources being devoted to making generative language models more useful. As a result, the market for detection software will likely be smaller for the time being. Additionally, Google has not yet released its generative model, so it is unclear whether the current detection efforts - which are all based on text generated by OpenAI - will work for Google's model, which is different.

Given these factors, I am skeptical that an enterprise like McGill could roll out an effective AI plagiarism detection software. Offensive capabilities in the realm of text generation will likely develop faster than defensive capabilities, and students may be wrongly disciplined based on probabilistic detection. Additionally, enterprise software licenses are expensive, and it may be a waste of resources to invest in detection software at this time. I believe that detection may only become feasible once text generation is built into something like Microsoft Office or Google Docs.

Part 3: Pedagogical Tool

Language models also hold considerable potential as both a pedagogical and a research tool.

Pedagogy

Here is one example of how a language model could be used as a pedagogical tool. First, I think that this gets at the traditional learning objectives of skim reading, identifying important elements of a paper, and focused reading. Second, it also teaches prompting skills. Third, it will teach students the limits of language models because they may often be wrong. Finally, students won’t be able to do all of these tasks with zero-shot prompting; they will have to break the text into chunks to properly process it.

1) Bring to class a sample of ABSTRACTS
    a) Ask the students to identify the key elements of the abstract manually - ie. RQ, hypothesis, key concepts, etc..
    b) Then, have the students prompt a language model to extract the key elements of the abstract.
    c) Compare and discuss the results
2) Bring to class some journal articles w/ only the INTRO and CONCLUSION
    a) Ask the students to try to write an abstract after reading these
    b) Then, using the same INTRO and CONCLUSION as inputs, have the students prompt a model to generate an abstract.

        c) Compare and discuss the results.

Research

Here are a few examples of how language models can be used as a research tool.

1. **Text extraction/annotation**: this involves writing an extraction prompt for a text dataset. Here is an example: I input the UN General Assembly floor speech for Bolivia in 2008. I wrote the following prompt: "Which sentences in this speech are related to inequality?" The model extracted sentences that were related to inequality. Note: the 1000-1500 word limit still applies
    - PROMPT: Which sentences in this speech are relevant to inequality?
    - BOLIVIA SPEECH TEXT: [FULL TEXT HERE]
    - OUTPUT: Relevant sentences related to inequality:
        - "The privatization of basic services and the policies imposed by the World Bank and the International Monetary Fund (IMF) did not solve anything for the majority of Bolivians."
        - "It is the fight between socialism and capitalism. That historic struggle is being resumed now, but I feel that these battles are uprisings and rebellions of peoples against an economic model — against the capitalist system. If we do not understand that capitalism is destroying humanity, then I am sure — and I hope that no one will be offended — that we will not solve the problems of life, the problems of the planet or the problems of humanity."
        - "I am not accusing anyone of being an accomplice of capitalism. From a communal and labour struggle to an electoral one, from organizer to President, I have become perfectly aware that capitalism is the worst enemy of humanity."
        - "We need to work together, we need to participate in the struggles of our peoples for equality. Only Governments and leaders working together with their peoples can find equality. When we have equality, there will be social peace throughout the world."

2. **Chain of reasoning prompting:** when prompting a model, you can also ask the model to justify its reasoning. For example, I input the UN General Assembly floor speech for Bolivia in 2008. I wrote the following prompt: "Extract 1 sentence that is relevant to inequality and explain why. The model selected a sentence and wrote up a short reasoning.
    - PROMPT: Extract 1 sentence from the speech that is relevant to inequality and explain why.
    - BOLIVIA SPEECH TEXT: [FULL TEXT HERE]
    - OUTPUT:
        - i. Sentence: "We need to work together, we need to participate in the struggles of our peoples for equality."
        - ii. This sentence is relevant to inequality because it emphasizes the importance of working together to fight against inequality and create a more equal society. It encourages people to interact with each other and participate in struggles to ensure that all people have access to the same opportunities and rights.

3. **Text generation: point v. counter-point:**
    - This is the abstract from the presentation:
        - **i.** *International humanitarian action is as important as ever as Red Cross, non-governmental, UN and state actors continue to respond to humanitarian need including for food, water, shelter, evacuation and medical care in places such as Ethiopia, Somalia and Ukraine. Yet understanding of the principles that govern such action remains uneven. Analyses of the humanitarian principles typically take as given that humanitarian action should be undertaken in accordance with the principles of humanity, neutrality, impartiality and independence. From this starting point, the focus is often on the principles' meanings and/or the challenges of applying them in practice. This article, by contrast, takes several steps back to address foundational but largely neglected questions about whether humanity, neutrality, impartiality and independence can accurately be designated 'the' humanitarian principles; about how these principles came to govern the humanitarian sector; about their legal character; and about their normative content and, more fundamentally, whether the principles can even have fixed meanings.*
        - **ii.** The author posed the following question during the presentation (paraphrase). I used this as a prompt for GPT-3:
            - Prompt 1):
                - *Is it possible for humanitarian actors to be neutral and not take sides in a conflict situation in order to take action?*
            - Response:
                - *Yes, it is possible for humanitarian actors to remain neutral and not take sides in a conflict situation in order to take action. Humanitarian actors, such as the Red Cross, are bound by international humanitarian law, which requires them to be impartial, independent, and neutral in order to provide assistance to those in need. Humanitarian actors must not take sides with any party or become involved in the political, military, or ideological aspects of a conflict. Rather, their role is to offer assistance such as food, water, and medical aid to those in need, regardless of their political or religious affiliation.*
            - Then, I used the **prompt 1** and the **response** to generate this counterpoint output:
                - *Humanitarian actors are not always capable of providing impartial assistance due to their reliance on government funding and support. This can lead to partisan and biased decisions that favor certain groups over others. The neutrality of humanitarian actors is often challenged in conflict situations because they may be perceived as taking sides with one particular group or faction. Humanitarian actors are limited in their ability to access certain areas due to security concerns and the complexity of negotiating*

access with multiple actors. Humanitarian actors are often subject to political pressure from governments and other powerful actors, which can lead to the manipulation of assistance. Humanitarian actors may be restricted by funding constraints and limited resources, which can limit their ability to respond to crises in a timely and effective manner.