ChatGPT, GPTs, and LLMs Survey Paper

Michael McKinsey

mckinsey@tamu.edu

April 30, 2023

ABSTRACT

This survey paper investigates the field of Large Language Models, mostly focusing on Generative Pre-trained Transformers specifically ChatGPT, as it is state-of-the-art for most generalizable generative AI tasks at the moment. Evaluated are many of the successes and criticisms of GPT models in the various problems scientists are trying to solve now. Some of the critical issues span across domains and professions. The biggest of which is the transformer models' lack of human reasoning, which prevents the models from recognizing and correcting their blunders. Errors then compound along the generative chain and a once seemingly elegant flow of thought devolves into nonsense. Of course, this is an extremely complex problem, as it is the difference between current AI and the future of Artificial General Intelligence.

1. Introduction

Generative Pre-trained Transformers (GPT)s are a class of Large Language Models (LLM)s used in artificial intelligence, taking the world by storm in the current decade. To understand how models like ChatGPT, LLaMA, or BERT are making such large scientific strides, we must understand the breakdown of how they work.

From Wikipedia, A Language model is "a probability distribution over sequences of words." These probabilities when used in the right context can be utilized to perform tasks, such as predicting the next likely word when generating a sentence. Transformers are just one architecture of language models, they can also be implemented with other architectures such as recurrent neural networks (RNN)s and Convolutional Neural Networks (CNN)s. The meaning of "large" in LLMs is an indication that the language model is trained on a substantial amount of data, a situation we can leverage with the current phenomenon of exponential data generation and collection.

As stated previously, transformers are a type of architecture, introduced in 2017, that heavily leveraged the concept of "attention" to gain unprecedented performance in language tasks [12]. "Generative" simply indicates that the model will generate some form of content based on the input, whether text or images. "Pre-trained" alludes to the purpose of why these models require a large amount of data. The models are subsequently trained on the corpus of information to base their understanding of how to generate answers. Drastically different data will result in drastically different model behavior. Therefore, GPT models can be summarized as follows: models capable of complex and sophisticated restructuring of input data, often emulating new inferences or ideas.

2. Effectiveness in the Sciences

Although GPT is strictly born from science, that doesn't automatically indicate that it will perform flawlessly at scientific tasks. As already indicated, LLMs are what they eat, and appropriate data must be provided to achieve correct results.

ChatGPT initially found success in mathematical datasets as the input problems were mostly elementary, but when provided with college-level problems consistently failed to provide accurate

solutions [1]. Although ChatGPT will almost always provide a convincing solution, the model has no guarantee of correctness unless it has already seen the exact problem before. Therefore, when provided with problems such as proofs, it fails to convert language into proper reasoning and often reverts to generating "nonsense" [1, 2, 3]. Some research even suggests that asking ChatGPT to reason results in worse performance. In one experiment with mathematical word problems, ChatGPT obtained a correct answer 84% of the time when providing an answer with no explanation and 20% of the time when required to show the steps and reasoning behind an answer [2]. Perhaps the chained architecture of the transformer model and the greater room for error makes it more statistically likely to make a mistake in one of the steps and carry that mistake into the following results. Even outside of mathematics in a different domain, theoretical physics, similar problems arise. When asked to provide definitions of well-known theories or equations, ChatGPT excels [3]. However, in lesser-studied areas, where there are fewer possible resources where the model could have been trained, it struggles. Such as in one experiment when asked to provide a list of "swampland conjectures", it provided "the bound on entropy conjecture", which *does not exist* [3]. Therefore, plenty of evidence indicates that blindly interpreting the outputs of LLMs as ground truth is unwise.

These cases show that ChatGPT and LLMs in general struggle to go beyond the surface of simply recalling information. They are largely incapable of identifying how to generate novel information from existing, and even if capable of doing so fail to ensure the correctness of an answer. Does this indicate that the sciences are too difficult for ChatGPT and AI to handle? The answer is no since it is already widely established that LLMs are satisfactory and in some cases even astounding at certain tasks, especially when involving recollection. ChatGPT achieves superhuman performance on the bar exam and can recall any mathematical or physical theorem on the spot, but it happens to struggle in tasks that are equally difficult for humans [11].

3. Identification and Classification Tasks

While we understand ChatGPT's state-of-the-art performance in Natural Language Processing tasks, it calls into question how well the model and LLMs in general perform in advanced language understanding. Such as genre identification, sentiment analysis, and standpoint extraction. Also, we contrast the generalized ChatGPT to other models that are specialized in each task.

Genre categorization requires a higher-level understanding of the text, a sort of summarization. Generalization becomes more difficult when text bodies consist of multiple different genres and less typical examples, so the model must have a solid understanding of the target. When compared to a specifically tuned genre classification model, X-Genre, ChatGPT performed slightly worse but still comparable on public datasets and performed superiorly on a held-out dataset [4]. This result is impressive since ChatGPT is not explicitly tuned for this task and illustrates the power of the generalizability of models trained on large, varied datasets. Greater generalizability is evident in the potential to solve tasks that the model did not face in training [4, 5]. When compared to RoBERTa (GPT), Word2Vec (NN), and BoW (baseline), ChatGPT performs state-of-the-art accuracy in sentiment analysis and almost SOTA in suicide tendency assessment [5]. RoBERTa, a GPT model, obtains the highest accuracy in the latter, so even when ChatGPT does not perform the best, it is still a GPT model which triumphs. This further demonstrates the capabilities of the transformer concept in general, no matter the backing implementation.

When trying to judge "for" or "against" in the domain of political commentary, ChatGPT is SOTA when compared to BERT, LSTM, CNN, and GCN models [6]. ChatGPT is also able to provide reasoning for its decisions, which is not a common feature among AI models [6]. This bonus can be explained in the sense that ChatGPT is designed specifically to hold dialogues, so follow-up questions on a decision are happily indulged by the agent. While not addressed in this last paper explicitly, one issue that degrades the accuracy of language models is the detection of sarcasm, as tone of speech becomes more important, which is a detail lost over text.

While it is clear that these language-related tasks are the most relevant next step for a language model to become an expert at, it is not evidently clear how to improve performance. As we are at the point where generalized models can beat out specialized models, it is suggested that the age of designing and tuning specialized models for specific tasks is widely over.

4. Humor Me

In this section, we take a closer look at how GPT models are impacting and creating new media. As we drive closer towards models that perform like humans, our models become more capable of creativity, and new interesting and entertaining results.

One study attempted to see if models could evaluate research paper titles and their subsequent abstracts based on if they were funny and how that impacted the quality rating of the subsequent paper. The experiment found that ChatGPT and BART (Transformer) models were able to perform the best, while the difference between the two was that BART was specialized to the task [7]. This is a difficult yet interesting task since humor and quality ratings are both subjective, but by collecting large amounts of data and comparing that with human judgments, we can somewhat approximate the comedic aptitude of LLMs. Another cultural-related task is the creation of dialogue agents for characters in a story. In the context of computers and media, this is an extremely interesting application of sophisticated AI models, as dynamic characters present a whole new immersive experience into a storyline.

A paper that addresses this task takes a specific introspective into a fictional universe by developing a Harry Potter Dataset in an attempt to create a legitimate Harry Potter dialogue agent [8]. Again, the implications of this, if it were successful, are huge in terms of media, since it opens the door to all sorts of unique and creative interactions between humans and computers. ChatGPT's specialization for dialogue also helps in this case since the task is bringing a character to life. However, this is a complicated task since many intricate factors go into predicting a character's actions in a specific situation and interpreting their state of mind [8]. It is not always deterministic what action a person will take based on their personality or character, and therefore the probabilistic nature of AI models is leveraged here to sometimes select different actions from the norm. It is still tricky to determine whether the character would've made such a choice, as the line between fiction and reality becomes blurry.

The commonalities between the two papers here are self-evident, as entertainment value and media generation will always be critical components of generative AI. As generative models progress, we will see static characters brought to life, through the power of language understanding and interpretation.

5. Summarization and Plagiarism

Potentially the biggest known feature of ChatGPT other than its proficiency in dialogue tasks is its ability to condense large amounts of text accurately. Believe it or not, this is not something that NLP models have been able to do aptly in the

past and GPT/LLM models are specifically good at summarization [9].

At least one reason for this is due to the "large" in LLM; the vast quantities of data ChatGPT is trained on allows it to have a vast knowledge base where it can pull from different sources to explain large texts in a unique, elegant fashion [9]. One definite shortfall here with ChatGPT is its tokenization limit. It requires large data to be provided in chunks or a reference of a specific document and hope that document was included in the model's training data. If it wasn't trained on that document, then the answer it generates will be completely inaccurate. Despite all its shortfalls, the easy accessibility and powerful summarization skills of ChatGPT also make it a proverbial forbidden fruit for all forms of cheating [9, 10]. Whilst work has already begun on attempting to detect plagiarism using ChatGPT, its very nature makes it difficult to detect as a writer, since it has been specifically trained to act like a human and provide unique responses. While plagiarism detectors for detecting copied text from other sources stay the same, detecting whether a text was *written* by ChatGPT and therefore "plagiarized" in the sense that it was not written by the human author is a different story. However, one interesting evaluation from this paper is that prompting ChatGPT in regards to if it had written a specific text resulted in a 92% success rate in one study [10]. This result indicates that there must be some method to the ChatGPT madness, most likely lying in speech patterns and delivery. The experiment does not mention whether the responses may have included variants of phrases such as "since I am an AI model, I cannot have an opinion…", which ChatGPT often will include at the forefront of any politically or socially sensitive statements. Comments like these are trivially easy to detect and would not be beneficial to include in the measure of accuracy, as they can easily be manually removed. With GPT-4, these saving-face statements are evermore present, as OpenAI attempts to avoid making a biased or opinionated model [11]. They aim to detect and avoid making any stances on sensitive subjects and topics that AI models do not yet have the proper reasoning capacity to deliberate.

Overall, ChatGPT is an extremely proficient model in the sense of summarization and is a new problem to be considered for plagiarism detectors. But this should not be an argument against the development of ChatGPT as there are way too many upsides especially looking towards the future.

6. GPT-4

At the time of this writing, GPT-4 has only been out for a couple of weeks and is only available in a premium version, ChatGPT Plus, accessible for 20 USD per month. Firstly, this new advancement was covered minimally in this survey paper, as although many of the studies are recent, there has not been enough time to do any experiments with GPT-4, so the latest assessments are with GPT-3.5. Secondly, it is disappointing that GPT-4 is hidden behind such a steep pay barrier, while it is understandable that there are operating costs for running the model as an easily-accessible service and its arguably the most competitive LLM out there, the question begs: why not make GPT-4 open-source? Or at least provide some form of the trained model itself, similar to the way Stable Diffusion is offered. It must be somewhat ironic that "OpenAI" is developing a closed-source product.

Digressing, GPT-4 has very promising features, offering new capabilities such as handling multi-language text [11]. GPT-4 aims to tackle a lot of safety concerns such as harmful content, proliferation of weapons, protecting privacy, and addressing cybersecurity concerns [11]. GPT-4's general strategy is to avoid providing an answer if

the agent detects a tricky subject. This will be a disappointment for the tinkerers and red teamers but seems like a logical step in the right direction. In addition, OpenAI warns against overreliance on the answers given by GPT-4, as it still doubles down on incorrect answers while sounding convincing [11]. Overall, it seems like OpenAI has spent a lot of time working on the marketability and accessibility of GPT since GPT-3.5 and are gearing up to make it a profitable product to sell to companies, devoid of unwanted behaviors.

7. Conclusion

Even while sometimes harshly analyzing ChatGPT and other generative LLMs, we should understand that generative AI is a blossoming field and will continue to make strides and advances in the coming years and decades. Instead of focusing on the shortcomings of LLMs we should strive to utilize them where they are capable, as additional tools in the belt of a productive individual. This survey paper for the most part addressed ChatGPT, since it's the hottest LLM right now and is subsequently the most studied and documented. It will hopefully not become too dominant though, as the past decade of competition has been great for rapid improvement in the field of artificial intelligence and machine learning.

References

1. Frieder, Simon, et al. "Mathematical Capabilities of Chatgpt." ArXiv.org, 31 Jan. 2023, https://arxiv.org/abs/2301.13867.
2. Shakarian, Paulo, et al. "An Independent Evaluation of CHATGPT on Mathematical Word Problems (MWP)." ArXiv.org, 28 Feb. 2023, https://arxiv.org/abs/2302.13814.
3. Lehnert, Kay. "Ai Insights into Theoretical Physics and the Swampland Program: A Journey through the Cosmos with Chatgpt." ArXiv.org, 10 Jan. 2023, https://arxiv.org/abs/2301.08155.
4. Kuzman, Taja, et al. "Chatgpt: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification." ArXiv.org, 8 Mar. 2023, https://arxiv.org/abs/2303.03953.
5. Amin, Mostafa M., et al. "Will Affective Computing Emerge from Foundation Models and General Ai? A First Evaluation on CHATGPT." ArXiv.org, 3 Mar. 2023, https://arxiv.org/abs/2303.03186.
6. Zhang, Bowen, et al. "How Would Stance Detection Techniques Evolve after the Launch of Chatgpt?" ArXiv.org, 10 Apr. 2023, https://arxiv.org/abs/2212.14548.
7. Chen, Yanran, and Steffen Eger. "Transformers Go for the Lols: Generating (Humourous) Titles from Scientific Abstracts End-to-End." ArXiv.org, 20 Dec. 2022, https://arxiv.org/abs/2212.10522.
8. Chen, Nuo, et al. "What Would Harry Say? Building Dialogue Agents for Characters in a Story." ArXiv.org, 19 Dec. 2022, https://arxiv.org/abs/2211.06869.
9. Yang, Xianjun, et al. "Exploring the Limits of Chatgpt for Query or Aspect-Based Text Summarization." ArXiv.org, 16 Feb. 2023, https://arxiv.org/abs/2302.08081.
10. Khalil, Mohammad, and Erkan Er. "Will CHATGPT Get You Caught? Rethinking of Plagiarism Detection." ArXiv.org, 8 Feb. 2023, https://arxiv.org/abs/2302.04335.
11. OpenAI. "GPT-4 Technical Report." ArXiv.org, 27 Mar. 2023, https://arxiv.org/abs/2303.08774.

12. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010.