**The Implications ChatGPT and Its Technicalities:**

Language models have been around for many years and are used for various applications, from chatbots to text autocorrect features. A famous example is Google’s Bidirectional Encoder Representations from Transformers (BERT) which is a popular model in the Natural Language Processing (NLP) space and has also been instrumental in improving the Google Search experience among other things. Recently, OpenAI released a research preview of their new language model, ChatGPT, and it took the internet by storm. It left people stunned by its capability to respond to different questions and prompts, have extended discussions, and even correct itself based on the user’s clarifications. It is no surprise that ChatGPT garnered over a million followers in the first week of its release. Less than a month before the release of ChatGPT, Meta also unveiled a large language model named Galactica which was specifically designed to aid in scientific tasks. However, Galactica’s reception was a far cry from the success of ChatGPT and it was shut down in only three days. This article will explore how large language models and ChatGPT specifically work, ChatGPT’s success, and its implications in our society.

ChatGPT is a large [language model](https://builtin.com/data-science/beginners-guide-language-models) and is different from conversational AI and chatbots that we often encounter, such as those on various customer service websites. Language models enable machines to understand language by determining the probabilities of the occurrence of potential sequences of words in a sentence based on previous words.1 Since both these models take in text input, give a response, and, in ChatGPT’s case, can also have follow-up discussions, you may wonder “How are they any different from a chatbot?” Let’s ask ChatGPT itself! I asked it if it is a chatbot and this is how it responded:

Essentially, it said that ChatGPT is not a chatbot because chatbots are more specific to different use cases and are more interactive, but its language understanding and text generation capabilities can be very helpful to create chatbot systems.

Before we discuss the success of this language model and contrast it with less successful models, let’s see how it works. OpenAI has multiple models including GPT-3 (Generative Pre-trained Transformer 3) and InstructGPT. Each has a different purpose: GPT-3 is trained on a much larger dataset and is meant for more general use in NLP applications, InstructGPT was created to follow instructions better and ChatGPT’s goal is to produce coherent and grammatically correct sentences. All of them use the [transformer neural architecture](https://towardsdatascience.com/transformer-neural-network-step-by-step-breakdown-of-the-beast-b3e096dc857f), which is a neural network architecture for NLP tasks. So, what’s the point of ChatGPT and InstructGPT when we already have the more general GPT-3? The original GPT-3 was trained on extremely large amounts of text data from the internet, which is the reason for its generality. However, this also causes it to be misaligned, i.e., the results it produces may be consistent but not necessarily correct in the required context.[1] Since GPT-3 generates probability distributions for sequences of words as described earlier, it lacks the context and background knowledge to consistently produce accurate output.¹ Misaligned models are not always helpful because they are likely to hallucinate (make up untrue facts), they may also produce information that is hard to interpret, and may even generate biased or toxic output.¹ Such errors are potentially dangerous. For instance, consider a situation where someone asks a language model how to treat a severe wound and the model produces untrue information, which can ultimately harm someone. Although the AI trainers at OpenAI have trained the model in a way that minimizes this behavior, this post by a Twitter user shows how one can trick ChatGPT to produce a potentially dangerous output.

To tackle the problem of misalignment, researchers employed both supervised learning and reinforcement learning, specifically the latter making ChatGPT unique.¹ Reinforcement learning is a technique where the model is “rewarded” for desired behavior, and penalized for unwanted behavior — similar to how people train their dogs by giving them treats when appropriate. They used a technique called Reinforcement Learning from Human Feedback (RLHF). This approach incorporates feedback from OpenAI’s human AI trainers into the training process in order to minimize false or harmful output.¹ An iteration of the training cycle involves three steps. The first is the supervised fine-tuning step where a prompt is taken from a prompt dataset. A labeler provides the desired output and this data is used to fine-tune GPT-3.5 (the language model on which ChatGPT is based) with supervised learning. In the second step, another prompt is given to the fine-tuned model and several outputs are generated. Labelers create rankings for these outputs which are then used to train the reward model. The third step involves using the reward model to further fine-tune the model obtained in the first step.

Large language models are not always successful, and ChatGPT is in fact an outlier. ChatGPT’s massive success occurred shortly after the failure of another large language model that was developed by a well-known tech giant. Whereas ChatGPT was fine-tuned to give coherent and conversational output, Meta’s Galactica was meant to be an assistant for scientific writing. Meta AI also published a paper about their creation of this model. The model was trained on a dataset of over “48 million papers, textbooks and lecture notes, millions of compounds and proteins, scientific websites, encyclopedias, and more.”³ Despite these efforts, the public demo of the Galactica had to be taken down in just 3 days due to harsh criticism. Like ChatGPT, Galactica also uses a transformer architecture, but clearly the fine-tuning was not sufficient despite the carefully curated dataset. The most common issue was that of bias and hallucination. Users complained that Galactica would make up information and present it confidently.⁴ This is dangerous and could potentially spread completely false information to users who are not well-versed in certain topics. Although Meta mentioned that large language models are prone to hallucination and should not be trusted completely, they also portrayed Galactica to be a practical research tool, which is why its tendency to hallucinate confidently was not well received.

With the introduction of ChatGPT, a large number of people have experienced the capabilities of artificial intelligence and have started to seriously consider the impact it may have on our society. Schoolteachers and university professors face challenges on how to deal with these new developments. ChatGPT is currently easily accessible and it is usually difficult to determine whether a submission was generated by AI. Especially in subjects like mathematics, we will see many students rely on this technology. Educators will likely have to embrace this new technology and form their teaching programs around them to maximize students’ learning.

The use of generative AI in scientific research is also an important issue, especially with the possibility of hallucination. There have already been cases where ChatGPT has been credited as an author in research papers.⁴ Research conferences and publishers would have to revise their policies regarding the usage of AI for writing papers or gathering scientific information.

ChatGPT’s success has made the impact of AI on our lives more apparent and calls for change in all spheres of life so that its use is regulated and fair.