Interacting with Large Language model Gemini and contrasting with Open Ai’s Chat GPT

By: Parker Christenson

When thinking about how advanced the large language models, various companies and many open-source models have been developed within the past two years. Ever since the success of Chat GPT 3.0, we have seen that models are able to help with various tasks that we must do throughout the day. Chat GPT has brought machine learning to the forefront of technology as it has become very advanced in architecture, and the detail of the responses that they are able to give. With, Google has also been making various pushes for the large language models as well. Google came out with Gemini. Gemini is also a very complex large language model that does not have any sourcing on the weights and biases much like Chat GPT. Considering that Chat GPT has a free version, and a paid for version, it makes sense not to release the fine-tuning mechanisms that help achieve the tension between the transformer architecture.

When using chat GPT you are given a very standard experience out of the box when creating a free account. The basic functions of the large language model are stored in an API, and the API is what is storing the model and giving the response back to the end user. This is the most common practice for LLM’s to be deployed. The response is fast, as it is just giving text back out, which is good because it is just being asked simple questions. I have noticed that there have been lots of other applications in which chat GPT is being deployed. For instance, on that I use almost every day when I am working is Co-pilot by GitHub. Which was purchased by Microsoft, and Open Ai allowed Microsoft to use Chat GPT for investment. When using co-pilot, the AI interface is very simple, as it is a extremely good code completion tool. I use co-pilot to help me write some of the more boiler plate code, and not so much the complex code. For example, making a mat plot chart is extremely easy, as it will help you with deciding your Y axis and your X-axis.

When using the Chat GPT interface though, there are lots of limitations on what the model can say and do. But when I was using Gemini, I found that the responses are a lot more inclusive of all possibilities. I am currently trying to build a computer that is specifically for making neural networks, and the limitations of Chat GPT are very noticeable. When I would ask questions about a specific SXM4 board connector which is almost a trade secret, Chat GPT would not answer it at all. But when using Gemini, the LLM was able to direct me to the part needed, along with modules I would install when downloading Ubuntu.

What I have been able to find about the models and the architecture has been somewhat hidden as they are in a very highly competitive space. Open AI has been somewhat public on they have been censoring or restricting prompts while google has too. Chat GPT’s architecture is based off of a large language model transformer, which relies very heavily on self-attention. Self-attention is a very complex way of having the model look back at what has been said and is able to then do some more inferencing when giving a response. GPT’s model design has been uploaded to YouTube by one of the leading engineers, but did not give key specifications to exactly replicate GPT4.

Here are some of the questions I asked the two models, to show the comparison.

Gemini

A screenshot of a computer

Description automatically generated

GPT-4

A screenshot of a computer

Description automatically generated

Gemini

A screenshot of a computer

Description automatically generated

GPT-4

A screenshot of a computer screen

Description automatically generated

Gemini

A screenshot of a computer program

Description automatically generated

GPT-4

A black screen with white text

Description automatically generated