The Age of AI Subscriptions

Ishan on 2023-03-24

TL;DR

TLDR: The author explains that while subscriptions are becoming the new trend in the tech world, the advertising model is what powered the world and made tech accessible to the masses. However, it is quite costly to run AI and even the best ad revenue is not enough to cover the costs of running an AI model. The author analyzes the costs associated with training the GPT-3 algorithm and concludes that it would cost around $20-30 million per month for Open AI to bear. Thus, charging users directly for the AI model they use through subscriptions seems to be the easiest way to make money and stay afloat.

Introduction

The rise of subscription-based models in the tech industry is hard to miss. From streaming services like Netflix to music apps like Spotify, paying a monthly fee for access to a service is the new norm. However, what if we could pay a monthly fee to access an AI? That's the future, and it's already here. In this article, we explore the costs of running an AI model like GPT-3, the challenges of generating revenue through ads, and the potential benefits of subscription-based models. We'll also examine whether charging users directly for AI services is a sustainable model for AI companies like OpenAI.

Advertising

Subscriptions seem to be the new trend in the tech world. From Netflix to Spotify, we are all paying a monthly fee to access a service. But what if we could pay a monthly fee to access an AI? This is the future of AI, and it is already here. This seems like the polar opposite of the advertising model of the 2010's. The advertising model is what made tech accessible to the masses.

No matter how much you curse the advertising model, it is what powers the world. The reason you smart phones costs so cheap: adverts, the reason your data costs just less, the reason you can access the world's most powerful search engine for free, all due to the power of adverts.

Now, adverts are really effective. Properly harvesting user data and properly monetizing based on that is hard and hats off to the marketing and strategy teams at the top companies who have been able to do it. Advertisements are huge money makers. Advertisements almost certainly produce results. That's the reason they still exist.

Costs of AI

Okay, for now, we have established that yes, there is a lot of money in advertisements. But is that money really enough to keep an AI company like open AI afloat?

Now before we get into it, I am sure you all must be thinking, well Google's algorithms and AI are top class in the world. If they have enough infrastructure to make it profitable, so would everyone else. Here, one must remember that though Google's recommendation algorithms are very complex, they are no where as big and complex as a text generator algorithm like GPT-3. The amount of resources, GPU and CPU power required to run a GPT model are quite costly.

So, let's do a research analysis on this. I don't exactly know how to write a technical report, so please manage with these finding of mine. If you find anything wrong with my finding, please feel free to reach out to me on twitter.com.

Now with that out of the way, let's get into the analysis.

References

These are very well written technical articles are if you are interested in the really technical parts of how it all works, I highly recommend you check them out or read the quick summary I have provided below

[1] The GPT-3 Economy: BDTalks

[2] Tom Goldstein on: twitter

[3] Demystifying GPT: Lamdalabs

Analysis

First up, it is straight up impossible to train the GPT-3's 175 billion parameters on a single V100 GPU. This what is traditionally used in data centers. V100 is a high performance GPU from Nvidia that is a huge. So how much time would it take to train a model on a single V100 GPU? 355 years and would cost a whooping $4.6million!

The appoarch open AI took is far more powerful.

Now, 175 billion parameters are just pieces of data that have to stored in bytes. Hence, they would need 175 billion x 4 = 700GB of free VRAM memory to work.

This is far more than any single GPU can offer. So, open AI uses a parallel system of GPU to get the work done.

> To train the larger models without running out of memory, the OpenAI team uses a mixture of model parallelism within each matrix multiply and model parallelism across the layers of the network. All models were trained on V100 GPU's on the part of a high-bandwidth cluster provided by Microsoft.

- LamdaLabs

Now, all of this is just process the parameters. Running the model's interface is a whole another story. Running all the tensor overflow models, numpy operators in themselves are very costly.

Overall, all of this ties in as a approximate of $0.0003 per word generated, which would be about because of the costs of running a A100 or V100 GPU on the azure cloud, costing $3 a hour. All this, and given that nearly 50million+ queries are made of chat GPT every month, this would land the cost some where to be 20-30 million a month for open AI to bear.

This is a huge cost. Sorry if I might have done any calculation errors or reference errors. Please let me know if I did something wrong.

Ad revenue || Subscriptions?

All this aside though, over all, it is quite costly to run AI. Now I don't have the patience to research the revenue in advertisements. But let's just say it is no where enough to run a good enough AI model.

Even if you set rate limits, token limits etc, your ad revenue will just now be enough.

So, obviously, the easiest way to make money and stay a float is my charging the user directly for the AI model they are using in the form of subscriptions. Hence, the user directly pays for accessing the AI and everyone is happy. But is it really that simple?

So, let's say I sign up to use an AI service, maybe one that generates articles for me. Now, I pay a subscription of $20, which is like the standard AI application fees now a days popularized by open AI's $20 for Chat GPT plus.

Now, through out the month, let's say I want a total of 10 articles, which is still a lot, but reasonable. Now, let each article have about 5000 words. So, let's assume each word costs $0.0003. So, each article would cost $0.0003 * 5000 = $1.5. And 10 articles would cost $15 dollars. So, what about the rest $5?

Solution Maybe?

You might argue that the sum total would not be much and it's true, but charging a subscription is not the best method. A better way would be charging the user directly for the tokens. So, the user pays for what they buy. Now, after this is done, you can then run ads on your service to make sure your company stays afloat.

I see no company doing this. Yes, open AI is sort of doing it with it's API, but it is more on the b2b end then actual user focused. I mean, doing so would but the power back in the hands of the user and not change would come to your company. You can maybe send them an annual bill, similar to how services like azure do.

So, if you do pitch this idea to your own company, you have my regard.

Should You pay?

I mean, all the services offering AI are charging you. Notion wants an upgrade, Canva wants an upgrade, hell I am sure Microsoft is in plans to charge for Bing AI too.

Now, for a company with a diverse portfolio like Microsoft, ad revenues might just be enough to make a little profit. The economics of which are not very clear just yet.

So, for now, I suggest, you hold your horses and wait for the major tech companies to figure out a way to monetize properly while using AI. Remember, for every new type of service, we always had a subscription model at first till the company figured out a better way to monetize. So, let's just give the Tech Companies some well deserved time and let's sit back and enjoy, the AI revolution.

Now, you might have noticed that this is a rather short article. This is because I am still learning AI technologies and I am not well versed in the technicalities of it. So, I am not able to write a very long article. But, I am sure that in the future, I will be able to write a much more detailed article on this topic. So, stay tuned for that. This being said though, all the information in this article is correct to my knowledge and limitation and I have done my research. So, if you have any questions, please feel free to ask me on twitter.

Thanks for reading this article. You can also find this article on my hashnode. If you liked this article, please consider sharing it with your friends and family. Also, if you have any questions, please feel free to ask me. I am always happy to help. All this aside though, I hope you have a great day and I will see you in the next article.

Ishan Writes 2023