ChatGPT for Enterprise Leaders
A Primer

2022 was a pivotal year for the field of AI, since Generative AI, a subset of AI, became startlingly good with the release of DALL-E 2, Midjourney, and Stable Diffusion during the early part of the year and with the release of ChatGPT towards the end of the year[1]. While Generative AI technology is still in its early stages of maturity with many concerns and limitations, it has reached a point where enterprise leaders need to understand the technology and consider its uses in their business strategy. This paper will pick one Generative AI product that has garnered our interest, ChatGPT. The technology behind ChatGPT is natural language processing (NLP). The NLP market is valued at 15B$ as of 2022 with penetration in BFSI, Healthcare, Retail, Automotive and High-tech[2].

WHAT'S INSIDE!

1. Explore the value of ChatGPT technology for your enterprise in the short and medium-term
2. Unpack the NLP technology in a business context with real world examples
3. Understand its limitations and plan better

Version 01 | March 2023 | a *lab45* Publication

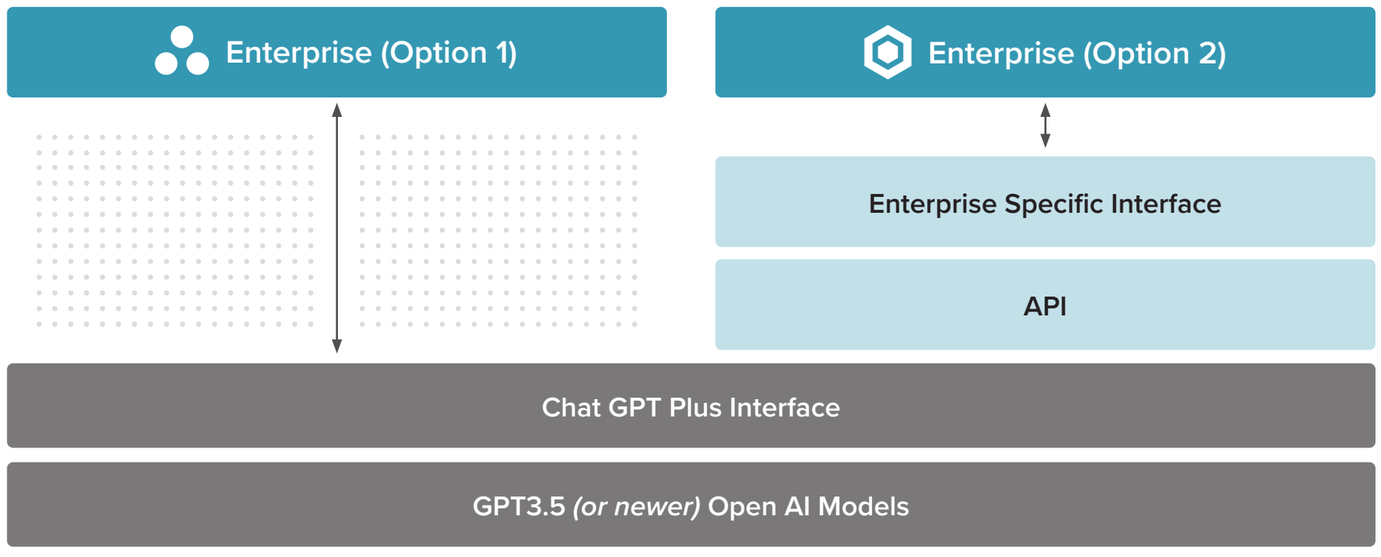
SECTION 01

The Value for you as an Enterprise Leader

Short-term (1-3months)

Deploy ChatGPT Plus[3] SaaS in selective functions within the enterprise and deploy ChatGPT API in some pilot programs to determine its ROI (functions and use cases are provided in the appendix).

ChatGPT is an extensively trained and massive AI model (175B parameters[4]) whose dataset is a significant portion of the internet until end of 2021. In the short term, it would be premature, expensive, and non-eco-friendly for enterprise leaders to invest in building this technology in-house. Further its limitations and ROI are not yet well understood. What would be prudent is for enterprise leaders to assign a small budget for different divisions of the company to leverage ChatGPT Plus, the SaaS version of ChatGPT, in their day-to-day operations (Option 1 in the drawing). The budget can also be used to do pilots with the ChatGPT API (Option 2 in the drawing). The divisions can then report back on the ROI and limitations.



Pros of Option 1

ChatGPT Plus has the potential to improve your employee throughput in a given function in the company. Employee throughput depends on their speed of learning and execution performance. ChatGPT can accelerate both as their learning and curation partner. A good portion of employee on-the-job learning is through the internet or books or word of mouth. All these forms of learning have a few limitations: they don't curate the learning and personalize it for a given situation, they don't have context and they do not have the ability to provide new ideas on demand. ChatGPT attempts to solve for all these problems in your employees learning. By providing ChatGPT the right questions, employees can accelerate their learning and execution process from days to hours. Example of this can be code developers or marketing teams. ChatGPT can generate code for fundamental use cases or help generate new content for marketing teams.

Pros of Option 2

The ChatGPT API[5] can be integrated in small pilot projects such as website chatbots, etc. to determine the user experience and feasibility and can also be used to create internal or external products/services which don't disclose confidential information. The benefit of this approach is that you can quickly assess the ROI for different function of your organization which will then inform your medium-term licensing strategy (in the next section).

1. https://openai.com/blog/chatgpt
2. https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933
3. https://openai.com/blog/chatgpt-plus
4. https://platform.openai.com/docs/model-index-for-researchers
5. https://openai.com/blog/introducing-chatgpt-and-whisper-apis

Considerations for Short-term

The limitation of the above approaches is that you need to put mechanisms in place to educate your employees to not divulge any confidential information since ChatGPT would learn that information. The outcome of this exercise is dependent on the competency of the individuals and their willingness to try new ways of working. Further limitations are discussed in the last section of the paper.

Medium-term (6months-1year)

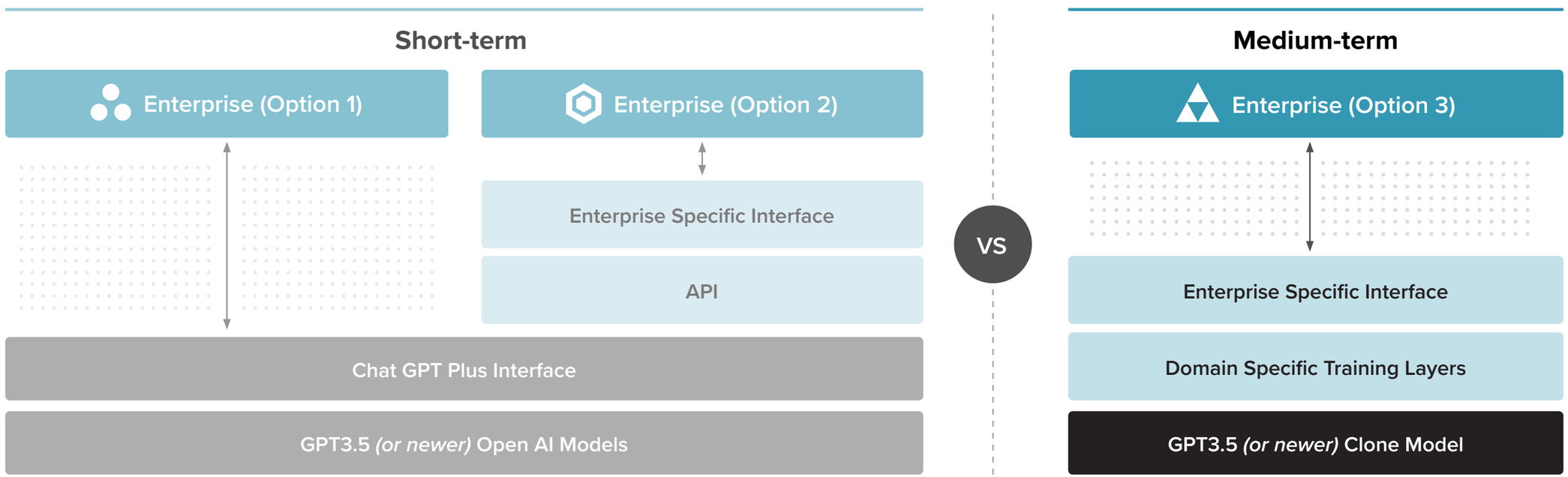
License GPT3.5, the base model of ChatGPT, and train it using your own domain specific intelligence. ***(Option 3 in the drawing)***

The SaaS model, ChatGPT Plus, is general access model and any information provided to it is not confidential and the information can be used by competitors. Further, the model is not trained on domain specific intelligence since it only has access to public data. To overcome the confidentiality issues and to make use of the full benefits of ChatGPT, you could consider obtaining a copy/license of the base model GPT3.5.

GPT3.5 can be trained by your company using domain/ company specific information (steps to train the base model and create differentiation are provided in the appendix).

Pros of Option 3

The pros of this approach are that you would develop a competitive edge being the first in the industry to combine the benefits of a powerful base AI model with domain specific intelligence that you have. This model can then be deployed to improve internal productivity or create a new set of products/services for your customers.



Considerations for Medium-term

1. ChatGPT is trained on GPT3.5 which is a newer version of GPT3. The full model of GPT3 (Davinci) is 175B parameters and translates to approximately 1TB of memory and requires a high-end GPU such as an NVDIA A100 and a high-end CPU such as Intel Xenon[6]. The basic estimated computing hardware cost to run the largest GPT model could be $25K-$50K[7]. Open AI has also released smaller versions of GPT3.5 such as Ada, Babbage and Curie which have fewer parameters.

2. If you choose to license GPT3 based models, the annual costs can be in hundreds of thousands of dollars per year depending on the usage[8].

3. There would be additional costs and energy expenditure associated with training the model on domain specific intelligence. These costs could start at 1M$[9] and could be higher depending on data availability, data quality, data formats, fine tuning the algorithm and may also require HFRL (Human Feedback Reinforcement Learning).

4. MS/Open AI needs to release a licensable model of GPT3.5.

Other considerations are discussed in the last section of the paper.

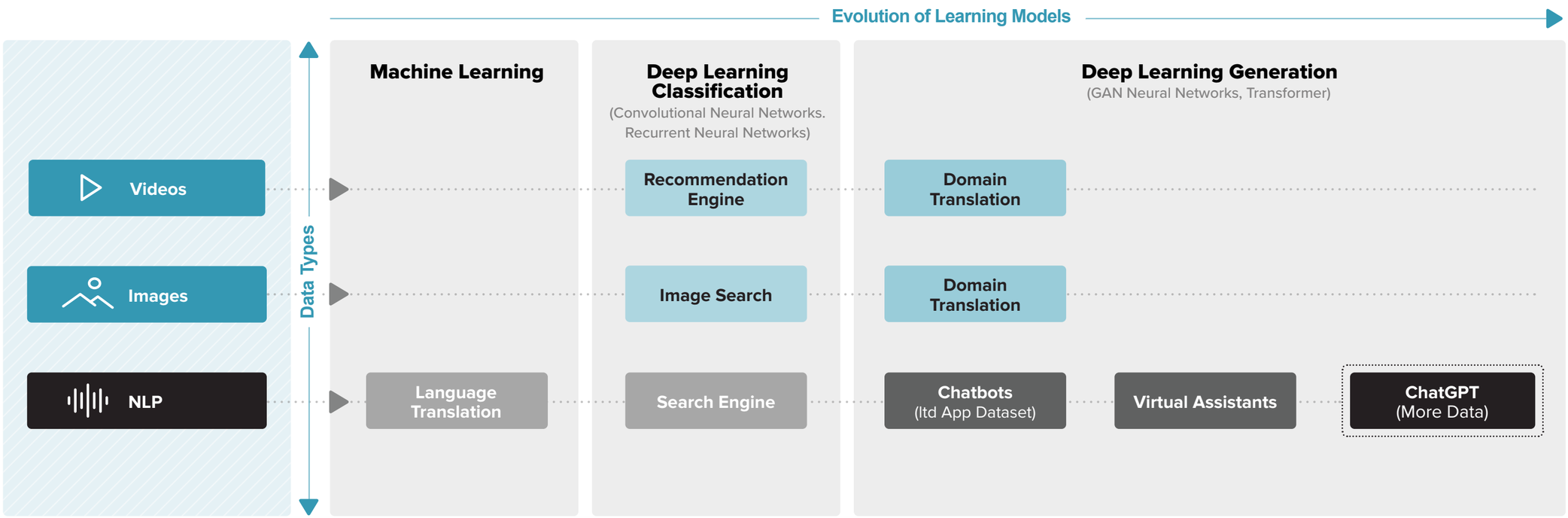
The steps in this approach would be to first develop a clear business/use case for needing this type of computational power. Second consider the smaller GPT3 versions, such as babbage[10], if possible. Third consider bringing in an expert who understands AI models and can develop a deployment and integration strategy.

> **The full model of GPT3 has 175B parameters. It translates to ~1TB of memory and requires a high-end GPU like NVDIA A100 & a high-end CPU like Intel Xenon.**

6. ChatGPT Query research
7. These numbers are approximate and only include computing cost. Further infrastructure and maintenance costs are not included.
8. https://techcrunch.com/2023/02/21/openai-foundry-will-let-customers-buy-dedicated-capacity-to-run-its-ai-models/
9. These numbers are approximate and can vary depending on the business use case, the compensation of data scientists, and other factors.
10. https://platform.openai.com/docs/model-index-for-researchers

SECTION 02

The Technology Behind ChatGPT (Natural Language Processing)

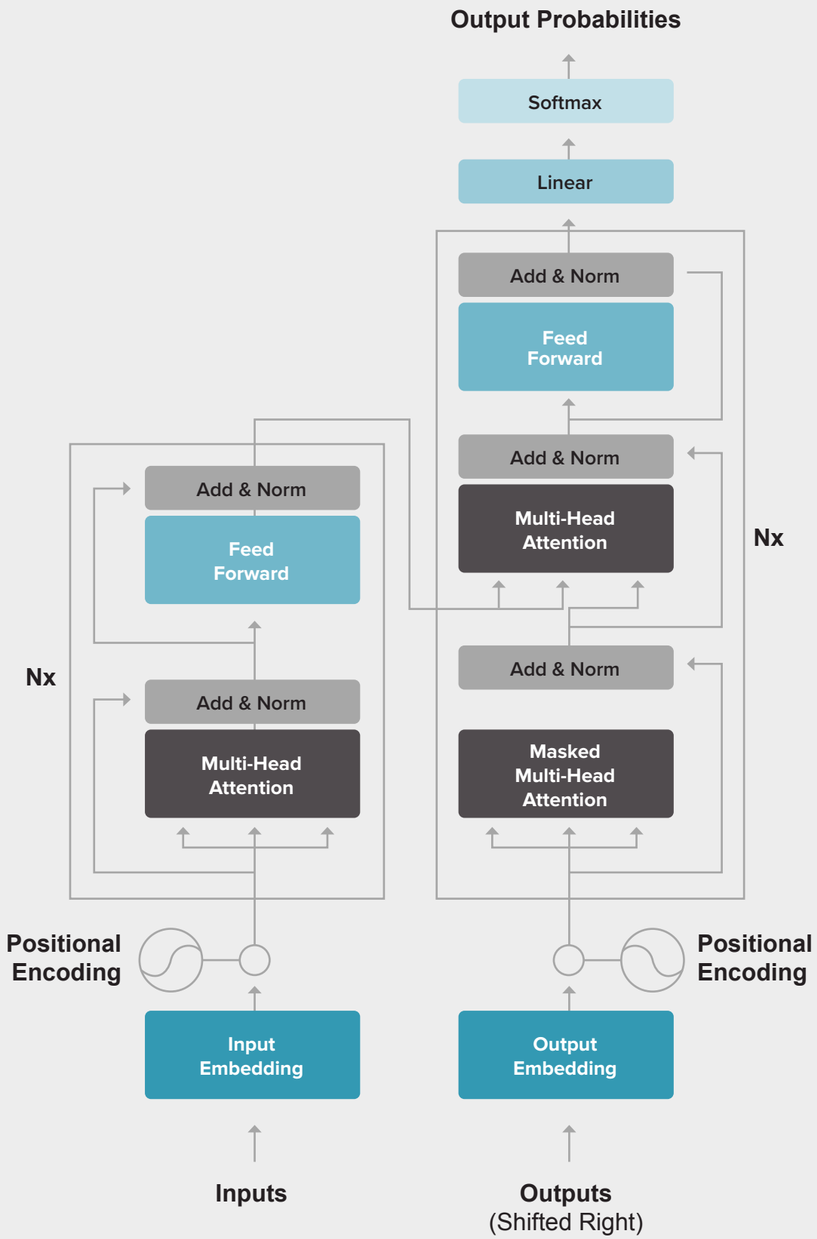


In this section, we will unpack the evolution of the technology behind ChatGPT with some business examples for easy understanding.

Machine Learning is a subset of the field of Artificial Intelligence, where computers learn from training examples to mimic human performance. The early stages of search engines, recommendation engines, etc. were based on machine learning. Today many of the predictive analytics integrated into business applications are based on basic machine learning models.

Deep Learning, an evolution of machine learning, uses a specific learning approach called neural networks to learn and requires more computational power. A deep neural network comprises of units or nodes called artificial neurons, which loosely model the neurons in a biological brain. They work on the principle of statistics assigning weights to data. Data moves through the network nodes based on weights or strength of connection between nodes. The early deep learning algorithms called Convolutional Neural Nets and Recurrent Neural Networks were capable of learning and classifying information. In our everyday life, this is seen everywhere today from chatbots to home assistants to recommendation and search engines.

The Transformer - Model Architecture



Source: Google Transformer Paper 2017

In 2017, Google first published a paper on **transformer models**[11], a type of deep learning neural network model that can not only learn from entire sentences but also generate text. Open AI, a Silicon Valley company, made a bet on this algorithm and invested significant amounts of money to build and train the transformer model calling it GPT. The transformer model works on the principle of an encoder and a decoder and uses the principles of attention and self-attention to learn from a large body of text and generate new text. The concept of attention is to understand the weights of each word in a sentence and decide on where to place importance. OpenAI has used this model to take the concept of Natural language processing to the next level and has been demonstrating its success over 4 generations from GPT 1 to 3.5. ChatGPT is the latest version of this neural network.

Some of the Factors Behind the Success of ChatGPT are...

1. **Massive Dataset:** It has been trained using sources such as common crawl[12] and Wikipedia to crawl a significant amount of the internet for data (499B tokens). (Tokens are pieces of words used for natural language processing (NLP). For text in English, 1 token is approximately 4 characters or 0.75 words)

2. **Massive Model:** 175B parameters vs its closest competitor, Turing-NLP at 15B parameters[13]

3. **Computational Power** (trained using Azure Supercomputers)

4. **Algorithm:** Open AI's significant investment into refining the transformer model for deep learning

5. **HFRL Training Methods:** Human Feedback Reinforced Learning.

11 / 12. https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

SECTION 03

The Limitations

ChatGPT's biggest strengths can also be its weaknesses. It has been trained on massive internet data, but it also knows only the internet (which as humans we know can be an inaccurate data source). It also lacks the ability to distinguish facts vs lies. So, this leads to some serious considerations.

Inaccuracy

- **Misinformation:** ChatGPT does not provide any references by default and can make confident mistakes (and sometimes factually / logically incorrect statements). It would be hard for a non-expert in a field to understand the blatant or subtle inaccuracies, leading to misinformation and related problems.

- **Obsolete World View:** ChatGPT may not be up to date with new information, especially around new concepts and connotations, and may provide incorrect confident yet inaccurate responses. For example, the original GPT-2 model will surely know what a pandemic is but will lack the detailed context around COVID-19 and its variants that has emerged in recent years.

- **Level of Context:** ChatGPT does not have the ability to understand emotions, or the hidden intent of the information provided to it. Therefore, it may not be able to provide accurate or appropriate responses to certain questions or prompts.

Bias

The training data that was used to train ChatGPT may contain biases, which could be reflected in its responses. This could lead to inaccurate or unfair results.

Maintenance Costs

ChatGPT is a complex system that requires regular updates and maintenance to ensure it continues to perform well. This can be a significant cost for businesses.

Legal issues

ChatGPT is a tool that generates text. It is important to be aware of the legal issues that might arise from using it, for example, copyright or trademark infringement.

> **ChatGPT has been trained on massive amounts of data from the internet, hence knows only the internet (which as humans we know can have inaccuracies and biases)**

Energy Usage and Costs

- Huge amount of energy costs for performing inference, training models and accumulating huge amounts of data.

- Data centers could have one of the most impactful energy usage on the environment.

- Single ChatGPT response costs at least 1 cent, it is estimated to cost $100K per million users (though cost will continue to go down).[13]

13. https://indianexpress.com/article/technology/tech-news-technology/chatgpt-interesting-things-to-know-8334991/

Appendix

ChatGPT Use Cases

Customers can use the following framework to determine where ChatGPT is applicable to their business.

- Situations where inaccuracy is not a challenge.
- Situations with no regulatory constraints
- Situations that do not involve divulging confidential company information.
- Situations with no real-time latency / throughput requirements
- Situations where human involvement is feasible to access the output.
- Situations that warrant the use of heavy computational power.

Building differentiated products using ChatGPT when everybody has access to the same underlying ChatGPT API

It is possible to build differentiated capabilities on ChatGPT by:

- Customizing the training data to suit their specific use case, such as using domain-specific language or incorporating industry-specific knowledge.
- Combining ChatGPT with other technologies or models to create a unique solution, such as integrating it with a task-specific model or using it as part of a larger system.
- Offering a unique user interface or experience for interacting with the model, such as creating a conversational application with a specific use case or target audience in mind.

To customize the training data for ChatGPT to suit a specific use case, you can follow these steps:

- Collect a large dataset of text that is relevant to your use case. This could include text from websites, books, articles, or other sources that are specific to your industry or domain.
- Preprocess the data to remove any irrelevant or sensitive information. This could include removing personally identifiable information, formatting the text in a specific way, or removing certain types of content
- Use the preprocessed data to fine-tune the pre-trained ChatGPT model on your specific dataset. This can be done using tools such as Hugging Face's Transformers library, which allows you to easily fine-tune a pre-trained model on your own dataset.
- Once the fine-tuning is complete, test the model on a holdout set to evaluate its performance. This can help you identify any areas where the model could be improved, and allow you to make adjustments as needed.
- Continuously monitor and fine-tune the model based on new data and feedbacks, to improve the model's performance over time.

Example Use Cases

Ad from product description
Turn a product description into ad copy.

ML/AI language model tutor
Bot that answers questions about language models

Translate programming languages
Translate from one programming language to another

Chat
Open ended conversation with an AI assistant.

Explain code
Explain a complicated piece of code.

Find more use cases
https://platform.openai.com/examples

Authors@lab45

Arvind Ravishunkar
Dinesh Chahlia
Nitin Narkhede
Noha El-Zehiry

Contributors

Aishwarya Gupta
Anindito De

wipro | Ambitions Realized.

lab45

Lab45 is a visionary space developing ground-breaking solutions to foster and accelerate ideation throughout Wipro.

At Lab45, engineers, research analysts, and scientists come together to infuse creative ways of incubating solutions for customers that will transform the future. It is a space filled with ambition at the vanguard of far-reaching research across cutting-edge technologies.

Established with the Silicon Valley culture of free-flowing creativity, Lab45's goal is to make bold ideas a reality and to invent the future of enterprise. So come, collaborate, and see what happens when ideas are left unbound.

Feedback

It will take less than a minute!

Click to Know More

Wipro Limited

Silicon Valley, USA

Bengaluru, India

wipro.com

Wipro's AI practice helps leaders maximize business value by integrating Artificial intelligence technology into their overall strategy and value stream. We offer a one-stop-shop approach where we help enterprises design and deploy AI across all its business functions. We leverage trustworthy AI that scales across business functions, is supported by actionable insights and is powered by curated data to help leaders build intelligent enterprises of the future. We further simplify AI infusion by using democratized methods, diverse and collaborative skill sets and by leveraging our partner ecosystem. Most importantly, we help you deploy AI in a responsible manner.

For more information, please write to us at aishwarya.gupta@wipro.com