embitel
A CARIAD GROUP COMPANY

Decoding **Transformer** Technologies : **A Comprehensive Exploration of** Llama 2 and OpenAI's GPT

Index

Introduction

The impact of Generative AI solutions such as GPT on our lives is profound. These technologies are transforming the way we interact with AI, acting as powerful facilitators of knowledge dissemination and manipulation.

From chatbots to virtual assistants, they have become integral to our daily lives, making information more accessible and tasks more efficient.In the wake of OpenAI's introduction of ChatGPT, there's a buzz around GPTs, but what exactly is a GPT?

It's a three-letter acronym, and each letter holds significant meaning. "G" stands for Generative, indicating its ability to produce content. "P" signifies Pretraining, highlighting the importance of training on past content to enable generation. Finally, "T" represents Transformer, the pivotal technology that ensures the generated content aligns with input and serves various applications. While both generation and pretraining are fundamental, it's the transformative power of the Transformer technology that truly drives the heart of this innovation. In this whitepaper, we'll explore the intricacies of GPT transformers, shedding light on their critical role in shaping the landscape of modern Generative AI solutions. In this whitepaper, we'll explore we will explore the intricacies of GPT transformers, shedding light on its intricate workings and its role in shaping the world of NLP.

Tracing the evolution of Transformer: the Driving Force Behind GPTs

Research over the years on how natural the Natural language processing (NLP) can be has brought us new computational benefits through technology. RNNs (Reinforcement Neural Networks) are a thing of the past with the advent of new-age computing power and technologies. Below we look forward to the new kid in the block that is transformers.

In the past, technologies processed input vectors sequentially, requiring labeled datasets and specific task-oriented encoder-decoders to consider semantics.

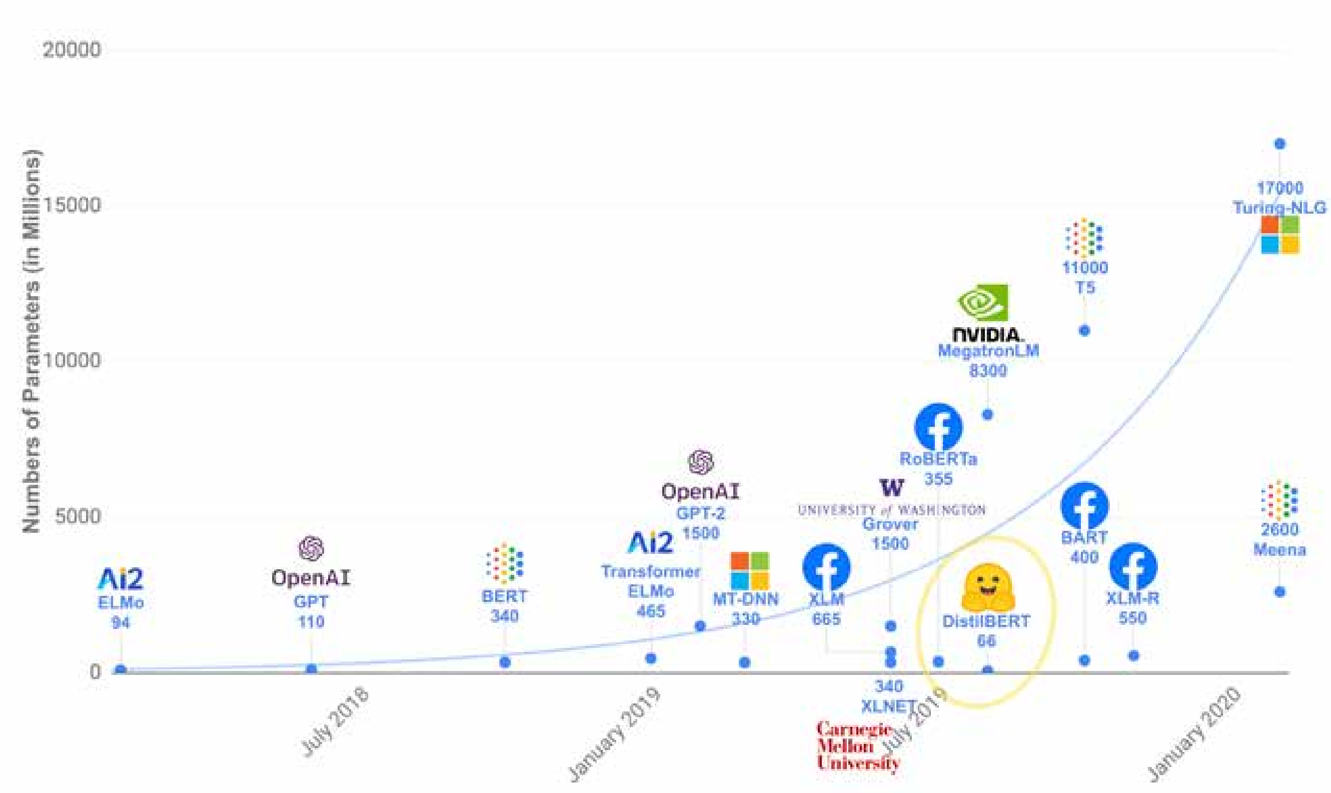
While they showed promise over the previous technologies but failed to a great extent in chatbots, handling long sequences, question-answer kind of scenarios and to top it all capturing the context. Capturing the context of the words in a sequence (sentence) and making a meaning out of their positional embeddings along with parallel processing of huge unlabeled datasets through large language models (LLMs) through multiple encoder-decoders parallelly processed are the key to all transformer technologies.

A sneak peek into large language models(LLM)

Large Language Models, like GPTs, represent a breakthrough in the realm of artificial intelligence. They are not sentient beings or repositories of hidden intelligence, but rather sophisticated tools that provide us with a "smart interface" to knowledge. LLMs are built upon artificial neural networks, primarily leveraging the Transformer architecture. What sets them apart is their training methodology, which involves self-supervised learning and semi-supervised learning. These models are capable of understanding and generating human language in a way that was previously unimaginable.

GPTs, in particular, have garnered immense attention for their ability to generate coherent and contextually relevant text. They are autoregressive language models, meaning they take an input text and predict the next token or word, making them exceptionally versatile. While earlier models required fine-tuning for specific tasks, larger models like GPT-3 have showcased a new approach called "prompt-engineering," allowing them to achieve remarkable results across various domains.

It all started with the publishing of a paper named “Attention Is All You Need” from Google Brains in the year 2017 and over a period of time, they have grown with more and more parameters with multimodal (text, image, voice, video etc.). Below pic shows the growth.



The evolution of transformers.
Image source:https://huggingface.co/learn/nlp-course/chapter1/4

Coming to the different types of transformers, there are broadly 3 types as below.

1. Encoder only – for purposes to understand the inputs and drawing a conclusion e.g., sentiment analysis.
2. Decoder only – for purposes of content is to be generated like story writing.
3. Encoder-Decoder (sequence to sequence models) – Content creation from input text.

While most of the open-source transformer models serve multiple use cases, GPT/transformers can be developed and be useful for single use cases as well for a specific domain e.g., CRM or Services where huge amounts of past data get captured.
Let's delve deeper into the fundamental operational principles of OpenAI's GPT (Generative Pre-trained Transformer) and Meta's T5 (Text-to-Text Transfer Transformer). We will also explore how each of these models operates.

Decoding Open AI & GPT APIs

OpenAI is an artificial intelligence research and development company focused on developing advanced AI models and technologies. OpenAI has been at the forefront of AI research and has released several groundbreaking models.

GPT (Generative Pre-trained Transformer) is a series of language models developed by OpenAI (GPT-2, GPT-3, GPT-4). These models are designed to understand and generate human-like text based on the input they receive. GPT models are built using a transformer architecture, which is effective for natural language understanding and generation.
GPT-4 is a newly launched model by Open AI on March 14, 2023. Which is trained with the 175B parameters and 45TB of dataset from different sources. It performs tasks including language translation, text generation, question-answering and more.

The OpenAI API is a service provided by OpenAI. We can integrate and use the API by initializing the API-key which we will get in Open AI. It provides a way to interact with models like GPT-3 or GPT-2. It enables us to generate human-like text with respect to the input given.

The API's mentioned below are the ones currently available from Open AI.

GPT-4 API: You need to have a subscription to use GPT-4

1. gpt-4: Updated version of GPT-3 and larger model size compared to the GPT-4, allowing it to handle complex tasks, and generate even more relevant contextual information.

2. GPT-4-0613: This version was last updated on June 13th, 2023. It can understand how functions are used but won't receive any more improvements or updates.

3. GPT-4-32k: It's like GPT-4 but can understand and use 32k character information at once. This helps it handle larger contexts in text.

4. GPT-4-32k-0613: This model is the combination of GPT-4-0613 and GPT-4-32k

5. GPT-4-0314 (Legacy): Last updated on March 14th, 2023, and it can understand how functions are used. But it's expected to become outdated by June 13th, 2024.

6. gpt-4-32k-0314 (Legacy): Last updated on 14th March 2023, with ability to recognize and process 32k characters at once. It's expected to become outdated by June 13th, 2024.

GPT-3.5 API: Free to use and available in https://platform.openai.com/

1. GPT-3.5 Turbo: This is a powerful AI model designed for chatting. The model is trained with the data up to September 2021.

2. GPT-3.5-Turbo-16k: It's like GPT-3.5 Turbo but can recognize and process 16k characters at once, which can be helpful for handling longer conversations.

3. GPT-3.5-Turbo-0613: This version was last updated in September 2023, and it's good at understanding how computer functions are used.

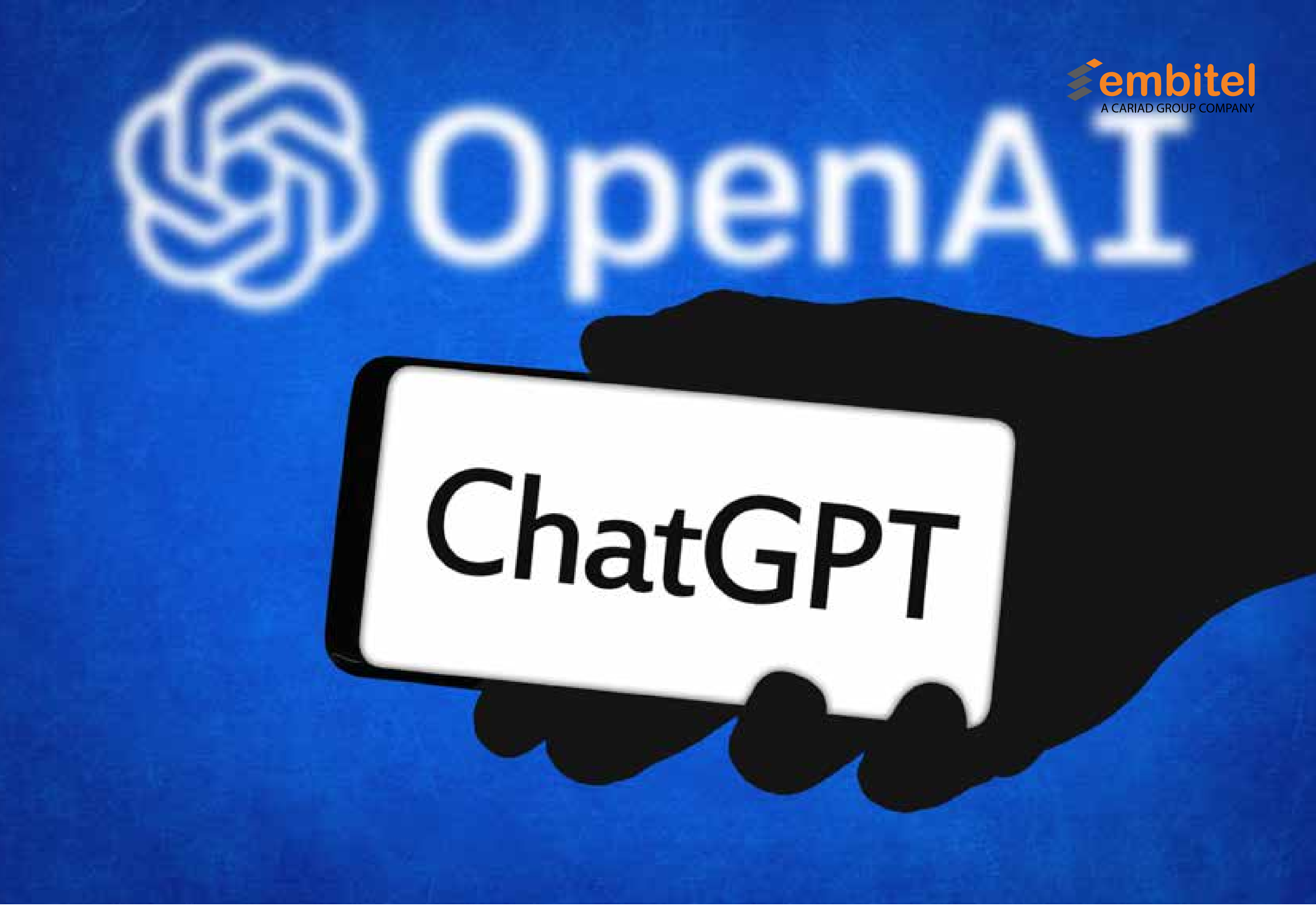
4. GPT-3.5-Turbo-16k-0613: This version is a combination of GPT-3.5-Turbo 16k and GPT-3.5 Turbo 0613

5. GPT-3.5 Turbo 0301 (Legacy): This version was trained with the data up to September 2021, and it will likely become outdated by June 13th, 2024, as newer versions become available.

6. Text-Davinci-003 (Legacy): This model is good for various language tasks and can provide longer responses.

7. Text-Davinci-002 (Legacy): Similar to Text-Davinci-003 but was trained a bit differently. It's also good at various language tasks and providing longer responses.

8. Code-Davinci-002 (Legacy): This model is specifically designed for tasks related to computer programming and code.



GPT Base API:

1.babbage-002: Replacement for GPT-3 ada and babbage base models with ability to process 16,384 characters.

2.davinci-002: Replacement for GPT-3 curie and davinci base models.
So, we will now look into GPT Transformers of OpenAI in more detail.

GPT (Generative Pre-trained Transformer):

- **Architecture:** GPT uses deep neural network architecture called Transformers. The Transformers models are built with the attention mechanisms and feed forward neural network.

- **Pre-training:** GPT models are trained with the 175B parameters and 45TB of text data. The models are trained to predict the next words in a sentence based on the context provided by the previous words. This process enables the model to understand language patterns and the context.

- **Fine-tuning:** After pre-training, GPT models can be fine-tuned on specific tasks, such as text generation, translation, QA assistance or sentiment analysis. We can adjust the model by using additional parameters while fine-tuning.

Attention Mechanism: The key innovation in the Transformer architecture is the attention mechanism, which plays an important role in understanding and processing the input data. Further, it can significantly speed up the training process of deep learning, in comparison to other models.

- **Autoregressive Generation:** It predicts one word at a time while conditioning on previously generated words. This process allows it to generate meaningful text.

- **Applications:** The Open AI Transformers are used in various applications like sentiment analysis, summarizing, question-answering, code generation and so on.

Now that we have delved into the concept of OpenAI, let's move towards understanding the intricacies of Meta AI.

Decoding Meta AI & Llama 2, **the next-generation LLM**

Meta AI is an artificial intelligence laboratory under Meta Platforms Inc. (formerly known as Facebook, Inc.). The primary focus of Meta AI is the development of various artificial intelligence technologies. One of Meta AI's recent projects includes Llama 2:

Exploring more about Llama 2

Llama 2 is Meta's open-source large language model (LLM). It is trained with more than 40% more data than Llama-1 and trained with two trillion characters. Capable of generating text and code in response to user query with respect to the system prompt given by the user.

System prompts are the prompts which are initialized to the Llama2 model to generate a specific result. For example,

system_prompt = “Given the provided information about a specific company, please respond to the question while maintaining a polite and respectful tone “

The above system prompts are given to the Llama 2 model to perform a Question-Answering task

Llama 2 is freely available and open source, allowing anyone to access and use the Llama2 models according to their needs. Basically, Llama2 provides a base and Chat model with the different numbers of trained parameters.

Llama2 is freely available, but you need to login to meta.ai to get any language models of Llama2. After successful login, you will get a mail from Meta with a url-key and link to git-repo. The url key will be active for 24 hours, and using that url key, you can download any model you want.

These are the available models

Llama-2-7b

Llama-2-7b-chat

Llama-2-13b

Llama-2-13b-chat
Llama-2-70b
Llama-2-70b-chat

You can use the meta-Llama2 model without downloading
You need to login to HuggingFace with the same email address to which you have logged in to Meta; after a while, the author of the Llama2 model will give access to you by confirming through mail.
From HuggingFace we can get fine-tuned llama models.

We will now be looking into Transformers of Meta AI

Meta’s T5 (Text-to-Text Transfer Transformer):

- **Architecture:** T5 is based on the Transformer architecture. However, it introduces a "text-to-text" framework, which frames all NLP tasks as a text-to-text problem.

- **Text-to-Text Framework:** In the text-to-text framework, both input and output are treated as text strings. For example, for a translation task, the input might be "Translate English to French: [English sentence]," and the output is the translated [French sentence].

- **Pre-training:** T5 is pre-trained on billions of parameters and large amounts of data. During pre-training, the model learns to convert input text into a target text, effectively learning how to map one text to another.

- **Fine-tuning:** Similar to GPT, T5 models can be fine-tuned on specific tasks by providing task-specific prompts and target text pairs. To fine tune on a specific task we need to use a custom dataset.

- **Applications:**We can use T5 Transformers for tasks like translation, summarization, question-answering, text classification, and more.

Below is the step-by-step representation of the Question-Answer model using Llama 2. The model is fine-tuned with the custom dataset and stored in a HuggingFace private repository.

Step 1: Set up logging

Step 2: Initialize documents using SimpleDirectoryReader

Step 3: Define system_prompt as " write your system prompt here"

Step 4: Define query_wrapper_prompt as "User query: {query_str}"

Step 5: Login to Hugging Face API

Step 6: Initialize a Large Language Model (LLM):

- Set context_window to 4096
- Set max_new_tokens to 256
- Set generate_kwargs with temperature=0.0 and do_sample=False
- Set system_prompt to system_prompt
- Set query_wrapper_prompt to query_wrapper_prompt
  - Set tokenizer_name to "insert your fine tuned model"
- Set model_name to "insert your tokenizer"
  - Set device_map to "auto"
- Set model_kwargs with torch_dtype=torch.float16 and load_in_8bit=True

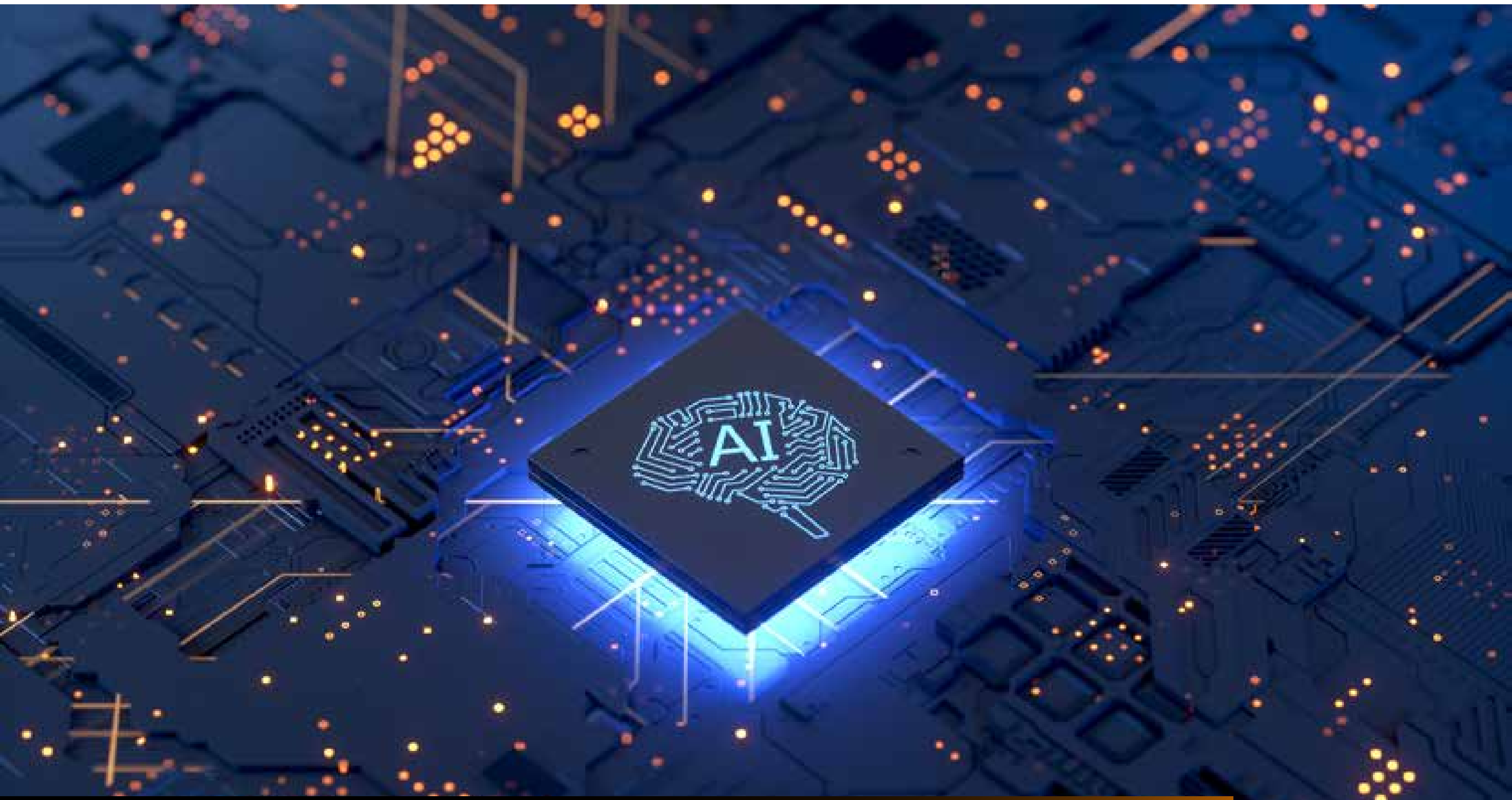
Step 7: Initialize an embedding model (embed_model) using HuggingFaceEmbeddings with model_name=”sentence-transformers/all-mp net-base-v2 “

Step 8: Initialize service_context with the following settings:
- chunk_size to 1024
- llm as the previously initialized LLM
- embed_model as the previously initialized embedding model

Step 9: Create an index (index) using VectorStoreIndex.from_documents with he loaded documents and service_context

Step 10: Create a query engine (query_engine) from the index

Step 11: Loop indefinitely:
- Read user input into the variable "query"
- If "query" is equal to "exit":
- Exit the loop
- Query the "query_engine" with the user's "query"
- Store the response in the variable "response"
- Print "response" to the standard output



Harnessing Llama 2:
A Game-Changer for Business Operations

Owing to its versatile natural language processing capabilities, Llama 2, Meta's Large Language Model (LLM), brings forth a wide range of business benefits.

Here are some key benefits that Llama 2 brings to the business landscape:

a. Enhanced Customer Support and Engagement

Integration of Llama 2 into chatbots and virtual assistants significantly enhances customer support operations. It facilitates immediate and accurate responses to customer queries, thus improving user experiences and continuous availability.

b. Data Analysis and Insights

Llama 2 can examine and summarize large amounts of text information. This is a valuable asset for businesses because it helps them gather valuable business and customer-centric insights into public sentiment, identify emerging trends, and make well-informed decisions grounded in data.

c. Multilingual Capabilities:

Llama 2's multilingual prowess enables businesses to expand their global reach and serve diverse markets with localized and culturally relevant content. This feature is invaluable for businesses with an international presence or those seeking to venture into international markets

d. Automated Code Generation with Llama 2

The model's ability to automate code generation offers a significant business benefit by streamlining and expediting software development. In addition to generating code, Llama 2 can assist in document creation and provide relevant solutions to programming challenges with respect to user queries.

e. Innovation

Organizations can harness the power of innovation by customizing the model, and to design solutions that precisely fit their requirements . With the ability to customize and fine-tune the model, businesses can create unique applications and accomplish specific tasks that set them apart in their respective industries.

f. Efficient Knowledge Sharing

Llama 2 can facilitate knowledge management and sharing within organizations by providing automated responses to internal queries.

Meta AI’s Llama 2 vs. Open AI GPT: **A Comparative Overview**

While both Llama 2 and OpenAI GPT models offer substantial advantages, businesses need to carefully weigh several important considerations when determining which model aligns best with their goals.

When evaluating Llama 2 and OpenAI GPT, several crucial factors come into play. Here's a comparative analysis of these models:

1. Customization and Open-Source Nature:

- Llama 2: Llama 2 is an open-source language model, that can be customized or fine-tuned according to specific organizational needs. This feature allows businesses to adapt the model to their unique tasks and applications, fostering innovation and tailored solutions.

- OpenAI GPT: OpenAI GPT models, such as GPT-3 and GPT-4, are not open source. However, we can access GPT-3 API’s without any costs involved, making them a convenient option for certain use cases.

2. Training Data and Scale:

- Llama 2: Llama 2 was trained on a vast amount of data, including over two trillion characters and one million new human annotations and also used 70 Billion parameters to train the model.

- OpenAI GPT: OpenAI's GPT models are also trained with a vast amount of data. The latest GPT-4 model trained with 45TB of data and 175 Billion parameters

3. Licensing and Costs:

- Llama 2: An open-source model, we can download the model by logging in to meta.ai

- OpenAI GPT: The GPT-4 model requires paid subscriptions per month.

To Conclude

To conclude, both Llama 2 and OpenAI GPT models are powerful tools with unique strengths. Llama 2's open-source nature and customization options make it an attractive choice for organizations to implement and use without any cost. On the other hand, OpenAI's GPT models offer key value- adds including - scalability and support for subscription models, and provide a high level of accuracy while generating text, making them suitable for a wide range of applications.

In the end, the choice between Llama 2 and OpenAI GPT models should depend on the specific needs of the organization upon the resources, cost, objectives and goals.

About the Authors

Jayanta Tewari
Head – Service Delivery, Operations & CSR

Jayanta is a seasoned professional in the dynamic landscape of the AI/ML-powered economy, renowned for his innovative and agile approach. With a remarkable career spanning over 22 years, he has left an indelible mark through the successful delivery of projects and programs across diverse domains, including Analytics, Digital transformation, and Cloud services. His expertise extends to pivotal roles in service delivery, pre-sales, and P&L management.

Shashank Ganesh

Shashank, a distinguished alumnus of NMAMIT College, holds a Master's degree in Computer Applications (MCA) and serves as our in-house expert on data science, artificial intelligence, and machine learning.
Shashank's remarkable eye for detail and flair for innovation are the driving forces behind his success.