Made By – Abhishek Kumar Singh

FALCON LLM ?

Introduction about the tool :

- Falcon LLM is a foundational large language model (LLM) with 40 billion parameters trained on one trillion tokens. TII has now released Falcon LLM – a 40B model.
- The model uses only 75 percent of GPT-3's training compute, 40 percent of Chinchilla's, and 80 percent of PaLM-62B's.

Process of its Development :

- Falcon was built using custom tooling and leverages a unique data pipeline that can extract high-quality content out of web data and use it for training a custom codebase, independent from the works of NVIDIA, Microsoft, or HuggingFace.
- Emphasizing data quality at a large scale was a key priority. LLMs have demonstrated a propensity for being influenced by the caliber of the data they're trained on. As a result, meticulous attention was dedicated to constructing a data pipeline capable of operating across tens of thousands of CPU cores, enabling swift data processing. This pipeline was engineered to extract top-notch content from the internet through a comprehensive process of filtering and eliminating duplicate information, ensuring the highest standards of quality.
- Falcon attains a level of performance equivalent to cutting-edge Large Language Models (LLMs) developed by prominent entities like DeepMind, Google, and Anthropic.

How Does it Work

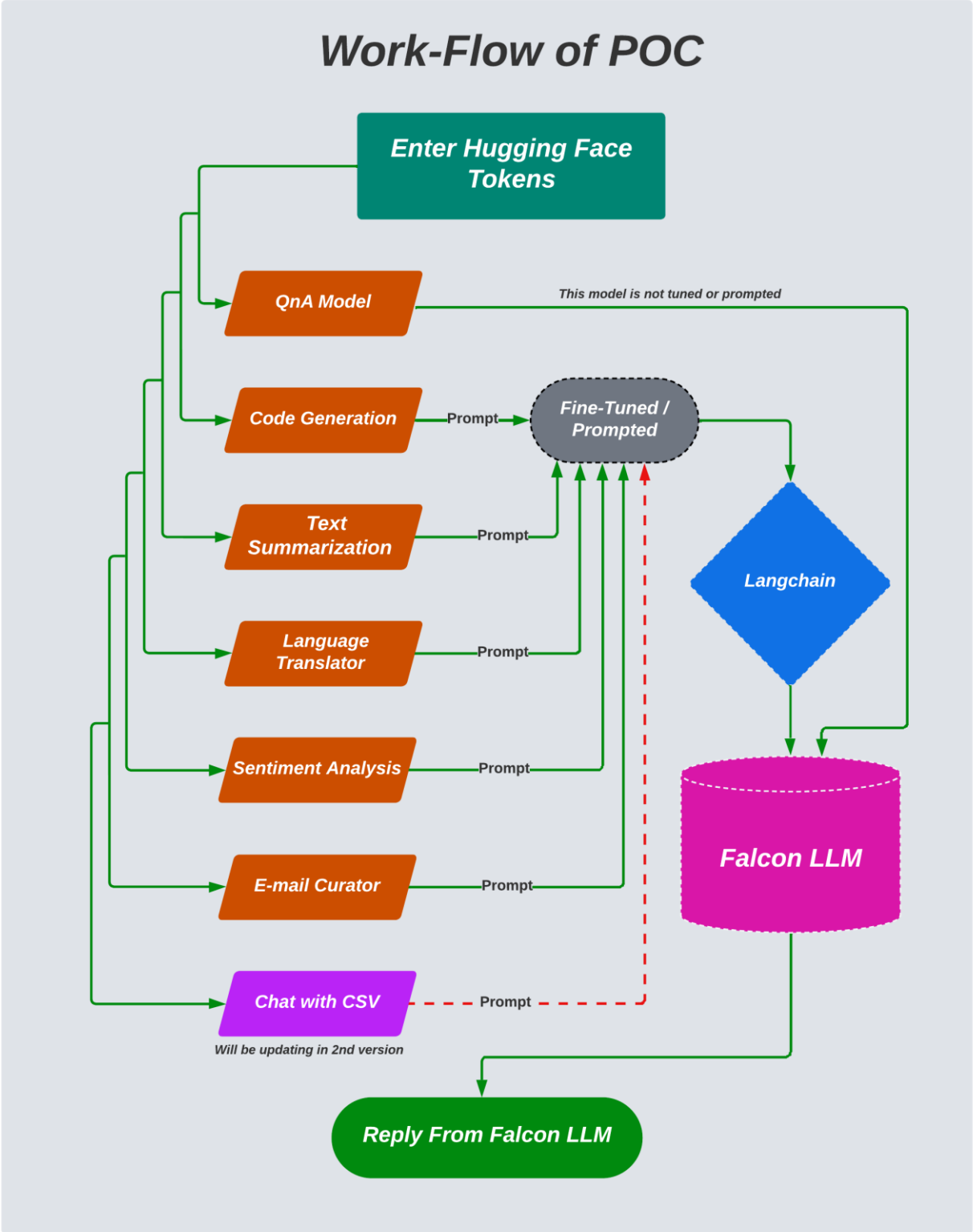
It is just like CHATGPT but it is open-sourced. Falcon is a 40 billion parameters autoregressive decoder-only model trained on 1 trillion tokens. It was trained on 384 GPUs on AWS over the course of two months. To broaden Falcon abilities, this dataset was then extended with a few curated sources such as research papers and conversations from social media.

How can we use Falcon?

- Generate creative text and solve complex problems.
- Used in chatbots, customer service operations, virtual assistants, language translation, content generation, and sentiment analysis.

- Broad use cases are foreseen by Falcon, although we are most excited about applications to reduce and automate “repetitive” work.
- Falcon will help Emirati companies and startups become more efficient, streamlining internal processes and giving back time for employees to focus on what matters.
- At an individual level, chatbots embedding Falcon will be able to assist users in their daily lives.

Workflow of the POC:



How to Use the Falcon Model:

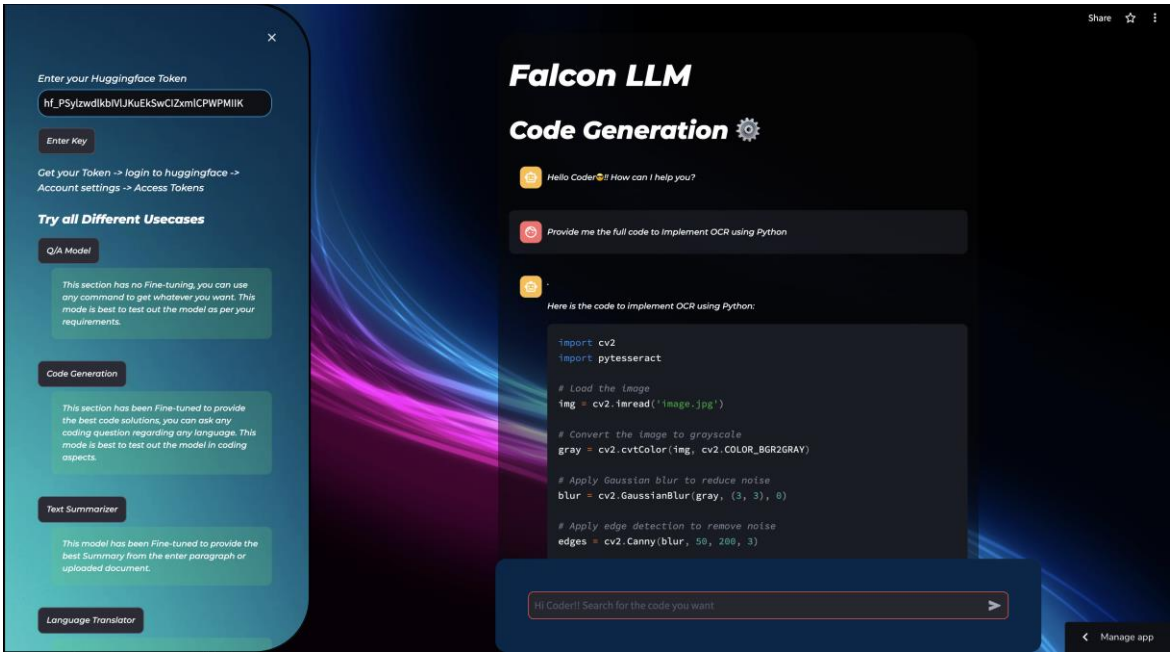
Go to https://falconllm.streamlit.app/ using Chrome (recommended for a better experience)

Currently we are using Falcon 7B model

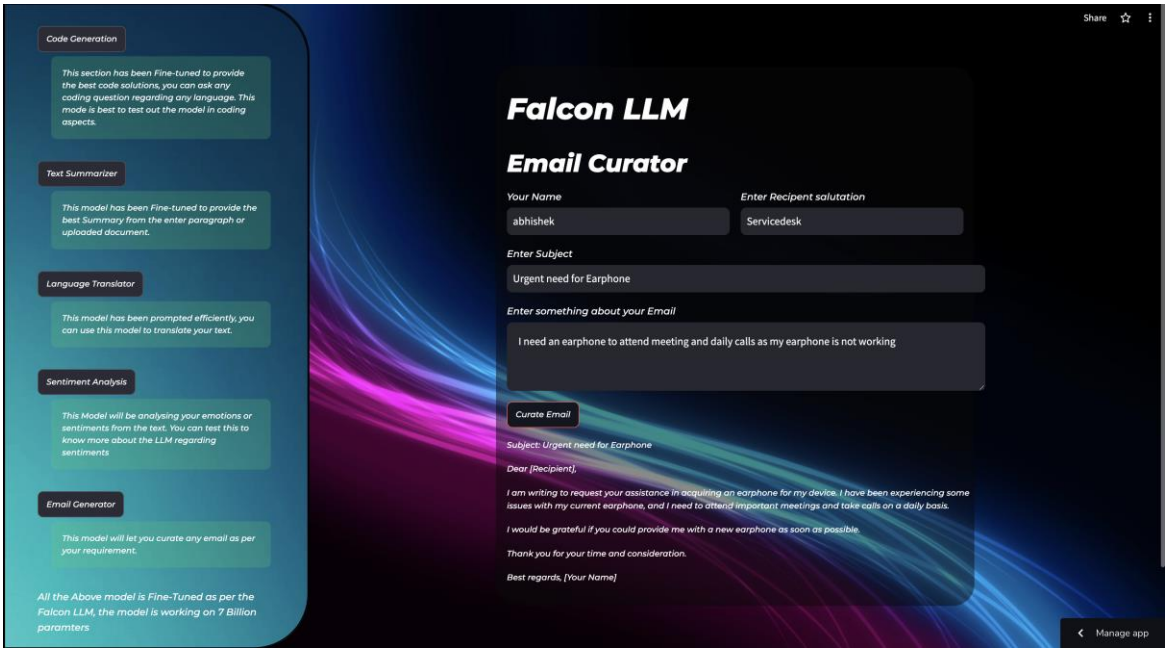
1. Enter your Hugging face Token to access the model.

2. After Entering the Token you can access all the model.

3. You can select any Model to use it . (Here is an example for the code generator)



4. Here is the Email Curator using FALCON



5. You can go ahead with Exploring FALCON.

Pros of FALCON:

- Falcon-7B and Falcon-40B along with their instruct versions are free to run on your own systems. You need 16GB of memory to run Falcon-7B whereas Falcon-40B required at least 100GB memory.
- The max Token for Falcon is 8196.
- It is fast. Falcon LLM can process text at a rate of 1000 tokens per second, which is much faster than other LLMs. This makes it suitable for real-time applications such as customer service chatbots.

- It is versatile. Falcon LLM can be used for a variety of tasks, including:
    - Natural language understanding
    - Natural language generation
    - Machine translation
    - Question answering
    - Text summarization
    - Code generation
- It is cost-effective. Falcon LLM is free to use for research and non-commercial purposes. This makes it a cost-effective option for businesses and organizations that want to use LLMs to improve their operations.

Limitations:

- It was trained on a dataset that was mostly in English. This means that it may not perform as well with other languages.
- It can be computationally expensive to use. This is because it is a very large model that requires a lot of processing power.

Conclusion:

Falcon LLM is accurate. It has been shown to be more accurate than other LLMs on a variety of tasks. For example, it has been shown to be better at generating text that is factually correct and not biased.

Overall, Falcon LLM is a powerful and versatile large language model with a wide range of applications. It is open source, large, fast, accurate, and versatile. These factors make it a valuable tool for researchers, developers, and businesses alike.