Fine tuning the LLaMa 2 model with custom data

The world began its journey on the path of digitization a few decades back. Today, it is walking a path that is immensely influenced by Artificial Intelligence (AI) and its advancements. AI has paved its way in every nook and corner of our lives. Some of the fields are:

- Business development
- Content production and management
- Coding
- Research
- Multimedia
- Design

Having said that, there is also an immense advancement in the usage of AI. Machine Learning (ML) is an application of AI where the machine is inferring mathematical models of data and learns without any direct instruction. This process of learning enables the machine to learn efficiently on its own based on experience.

What is LLM?

A foundational model in ML is the Large Language Model (LLM) which uses Deep Learning (DL) and large data to understand language and also generate new content very similar to content generated by humans. These models are trained extensively on various types of data such as texts, images, and data from books and websites to ensure Natural Language Processing (NLP).

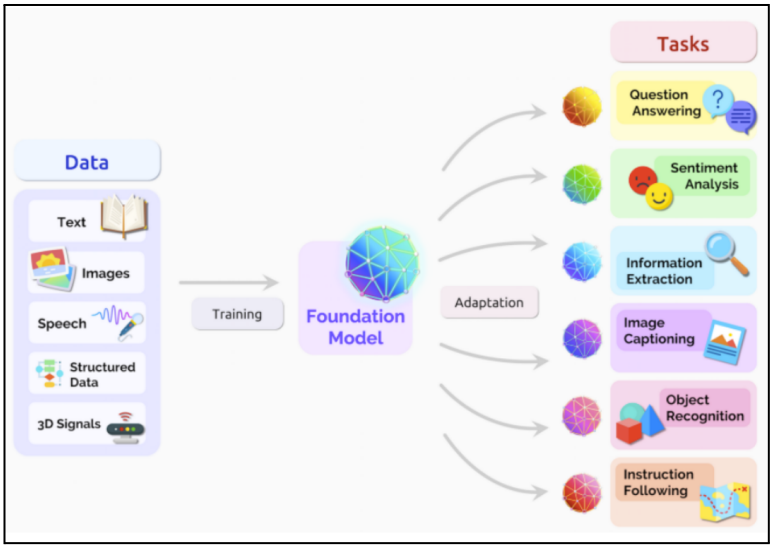

Figure 1: Foundational Model, Source: ArXiv

When users provide a prompt or ask a question, the LLM responds accordingly in no time. This reduces manual work thereby giving a good amount of scope to automate processes. It also saves money and time.

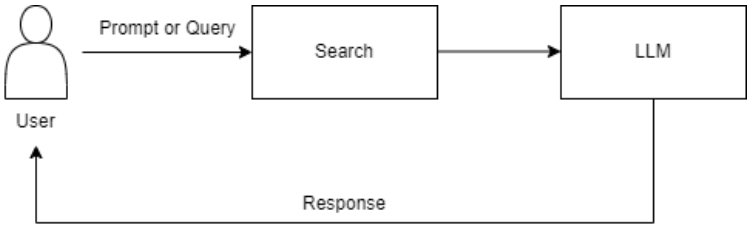

Figure 2: LLM workflow

What is LLaMa?

LLaMa stands for Large Language Model Meta Artificial Intelligence. It is provided by Meta AI. This language model is rigorously trained with large volumes of data spanning from 7 billion to 65 billion parameters. It is said that this language model is smaller and can be retrained and fine-tuned as per needs. It is trained using the top 20 languages having the highest number of speakers across the world. This language model receives a set of words as input and anticipates the next word and thus produces a wholesome text.

What is LLaMa 2?

LLaMa 2 is the LLM brought by Meta AI and Microsoft together. These tech giants have come together and introduced LLaMa 2 intending to make AI available and easily accessible to all. This model is available for research and commercial use, unlike its previous version. This model helps developers across various organizations to develop generative AI tools which in turn can be used to create new content including audio, code, images, texts, simulations, and videos.

What is LLaMa 2 fine-tuning?

Fine-tuning, as the word suggests, is making adjustments to obtain better or desired performance. The language model, LLaMa 2 can also be fine-tuned by adjusting and adapting this pre-trained model to perform specific tasks. In this process, smaller and targeted custom datasets are used to train the LLaMa 2 language model. Initially, every language model is pre-trained with large amounts of data which helps the model to learn general aspects of languages. Fine-tuning utilizes this aspect and ensures better performance and an in-depth understanding of a specific domain.

Eg: LLaMa 2 can be fine-tuned to identify symptoms of diseases in a medical text or predict stock prices based on financial news.

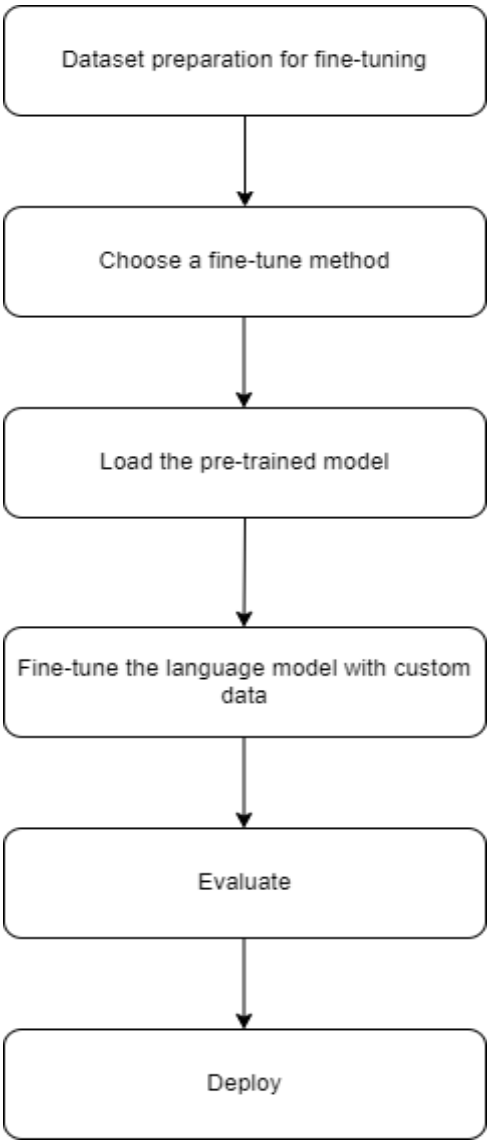

Figure 3: Fine-tuning process

Why fine-tune LLaMa 2?

The gains of fine-tuning a language model like LLaMa 2 are limitless. Some of the prime reasons are:

- Customization
- Data sensitivity and compliance
- Domain-specific language
- Enhanced performance
- Improved user experience

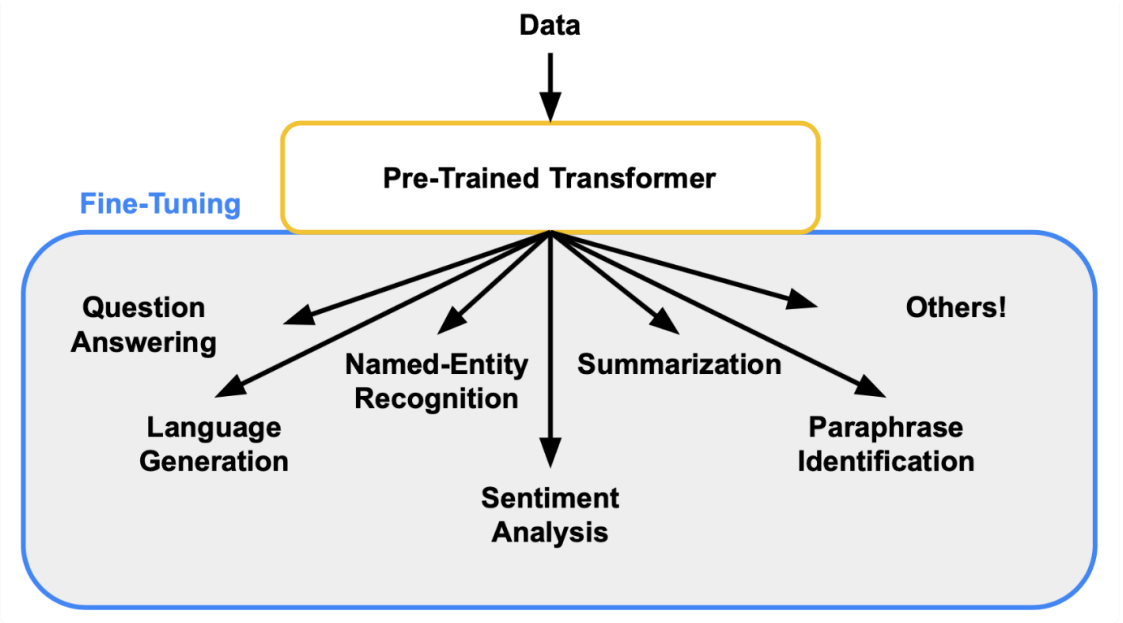

Figure 4: Benefits of fine-tuning, Source: AssemblyAI

Fine-tuning LLaMa 2 is highly beneficial for organizations as they already have a pre-trained model and they have to work on enhancing it to cater to their specific business needs. This reduces development cost, operation cost, and product evolution time and cost, thus fast tracking the product deployment to market. The organizations can majorly focus on building their datasets for fine-tuning and need not worry about building the product from scratch.