The Process for Developing an LLM

Large Language Models (LLM) are artificial intelligence systems designed to understand, process, and generate complex ideas and sentences. They have been increasingly crucial in recent years as AI chatbots become integrated into various technologies. As these models become more embedded in our daily lives, it becomes essential to comprehend how they are developed to better understand their abilities, limitations, and ethical concerns.

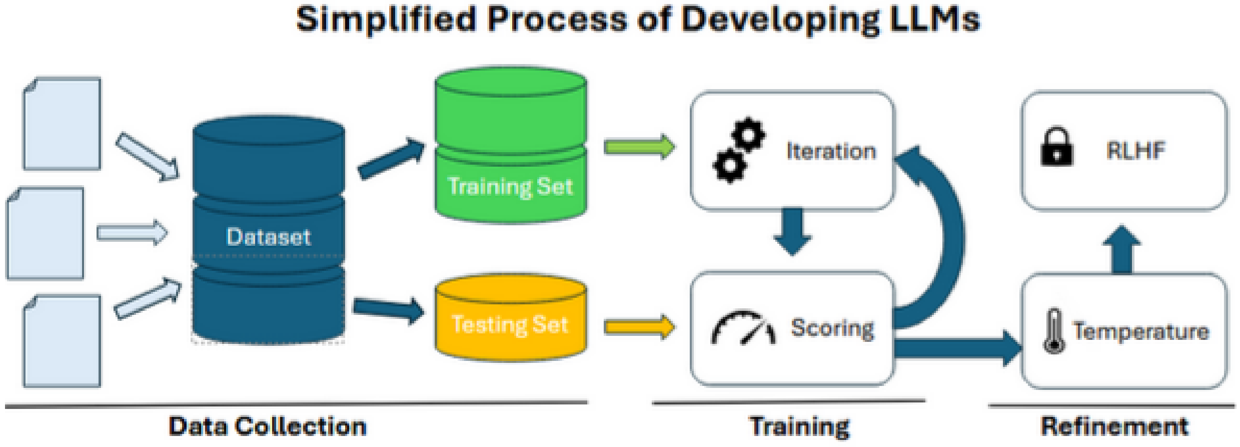

Figure 1: An overview of the LLM development process

Data Collection

Locating Datasets

The core function of all AI models' is to replicate patterns found in data, similar to how a student would study previous exams to prepare for a final. Unlike smaller statistical models that may try to predict a few points, LLMs require massive datasets to learn from as they try to capture intricate patterns found in human language, including ideas, questions, and responses. To generate these large datasets, researchers rely on a technique for data collection called *scrapping*, where automated software collects sentences through scientific journals, online books, websites, and more – often not intended to be used for training models.

An aspect of locating data that is often overlooked is whether it is sourced responsibly. This is an increasingly common ethical concern, as data scientists pursue increasingly larger pools of data to enhance models. While scrapping, unintentional infractions of copyright can occur, as scrappers often do not ask permission due to the sheer scale of the operation. While this step may seem small, these infractions can hinder the final model's use as this cannot be selectively removed from a model after training. Once a sufficiently large and ethical-sourced dataset is compiled, researchers continue to the next step.

Preliminary Analysis and Data Cleaning

A popular saying in data science is “Garbage In, Garbage Out”, emphasizing that flawed data leads to poor models. Regardless of how advanced the techniques used are for building a model, the quality of the final model is directly related to the quality of the data it learns from. Before training, researchers conduct a rigorous analysis of the dataset – identifying and removing errors, biases, and inconsistencies. Cleaning the dataset also serves as a checkpoint to determine if a model is even plausible, and that the data is meaningful and large enough to build a model of reasonable strength. After analysis, the data is prepared for the training phase.

Training and Evaluation

Model Architecture

Large Language Models, along with most machine learning models, draw inspiration from nature's strongest language processor: the human brain. Structurally, an LLM consists of many artificial *neurons* – mathematical functions that take in tokens (usually words or phrases) and outputs a prediction for the next word. While an individual neuron is simple, millions of neurons in interconnected layers can derive complex networks capable of understanding linguistics.

Models achieve this by giving each new output word a confidence score, indicating how likely the word is to come next in the sentence. This process, called *autoregressive generation*[2]*,* allows models to build sentences iteratively, predicting one word at a time, and adjusting its predictions based on prior outputs. Many LLMs use autoregressive generation since they can develop unique ideas while adhering to rigid sentence structures. The most significant issue with this structure is its *black-box* nature – a process that is so complex that its decisions cannot be backtracked. This raises ethical concerns about an LLM's ability to be accountable and the transparency of these models. Despite these risks, this represents the foundational architecture of LLM which is critical in the next step: training.

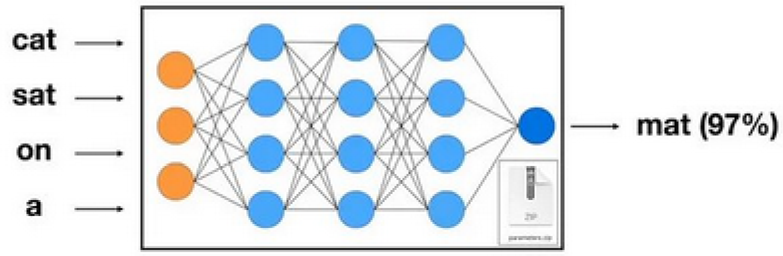

Figure 2: a simple representation of a LLM taking in the previous four words to predict the next with a level of confidence

Training Model

During training, researchers assess an LLM’s performance by evaluating it with training data – a dataset used to score the model's accuracy. Based on how well the model does compared to the expected result (think about an exam being scored against an answer sheet), and are provided an accuracy score which is used to determine the strength of a model.

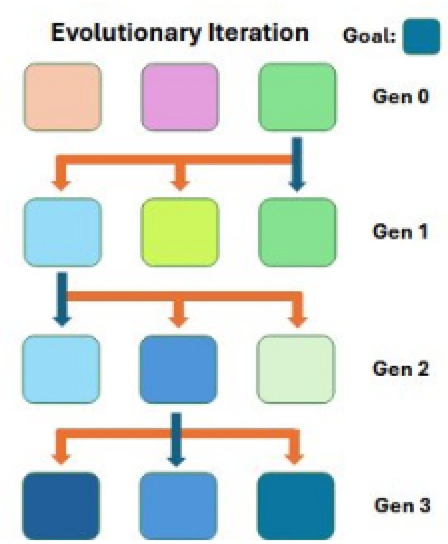

Figure 3: an evolutionary process where a random starting array become closer to a goal through generational randomness.

The goal of LLM training is to make models as strong, and thus as useful, as possible. One common approach to making stronger models is using a process inspired by nature: *evolutionary learning*[3]. This involves generating many of these models, all of which are provided random neurons, and are efficiently randomly guessing. However, some of them end up guessing slightly better than randomly, just by chance. The best of these models are then duplicated and have some of their neurons randomly tweaked. Similar to natural selection, each *generation* has some models that randomly get stronger, and the population gradually moves towards becoming a meaningful model. After many iterations, the model is strong enough for the final phase: refinement.

Refinement

Temperature

One of the defining characteristics of an LLM is its *temperature*, a unit used to describe how “creative” a model's generation is. On a technical level, this controls the randomness of the prediction process, and how willing the model is to select the n-th best option for the next word. The warmer a model is, the more likely it will choose less common phrases while a cooler model will be more likely to pick common patterns. Finding a balance between these is important, at its most extremes, a cold model is only mimicking text from its dataset, and a hot model will become so “creative” that it generates incoherent text. The desired outcome is to generate unique content while still adhering to the rigid rules of linguistics.

Another issue that can arise is model *confabulation* – when a model is hot enough to generate plausible-sounding “hallucinations” but is unable to fact-check itself. Though these errors will always exist, they can be mitigated by coercing the model through fine-tuning the temperature.

RLHF

Reinforcement Learning from Human Feedback (RLHF) describes the process of teaching or hardcoding a model to adhere to human ethics. This can range from preventing a model from making explicit jokes to something as extreme as preventing a model from generating instructions for illicit activity. The most effective route in preventing this is using a *feedback loop* – where developers manually judge the model's output. Understanding users' psychology and building robust RLHFs is an important final step before an LLM is ready for public use.

Conclusion

The development of Large Language Models is a meticulous process, encompassing data collection, model training, and refinement. Each phase is crucial and in conjunction, creates AI which revolutionizes how we interact in the world. Though this process is rigorous, there is still room to improve the ethics and insight into LLM, and understanding these limitations allows users to be more aware as society becomes more ingrained with these models.

References:

https://medium.com/@codersama/fine-tuning-mistral-7b-in-google-colab-with-qlora-complete-guide-60e12d437cca

https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f

https://www.ibm.com/think/topics/large-language-models