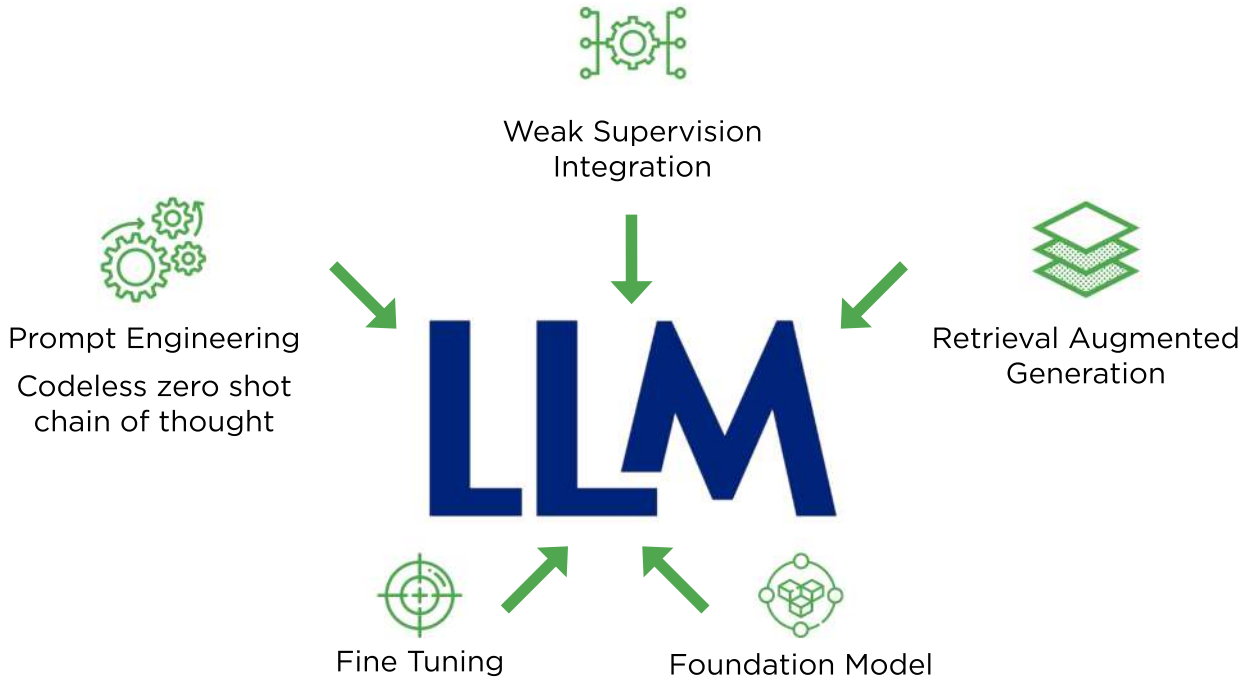
TRUTH OR HALLUCINATIONS IN THE AGE OF LLM

ChatGPT from OpenAI has taken the world by storm. Business leaders, from CEOs to VCs are foreseeing a tsunami of innovation surrounding the capabilities of Large Language Models (LLM), with wide ranging implications for customer support, healthcare and life sciences, finance and banking, human resources, supply chain logistics, legal / compliance, and even code development.

To put LLM into practice into your specific business operations, there are numerous considerations to be taken into account. For folks in healthcare, the obvious top concern is of course patient privacy. ChatGPT is the first to warn users that submission of any private information can lead to serious privacy and security risks. This leads to a swift response from competitive offerings such as Facebook LLaMa, Stanford Alpaca, Google Med Palm 2, Hugging Face Falcon 4B, which offer smaller footprints alternatives making it possible to operate on-prem instances of LLM. Google, on the other hand offers cloud-based Med Palm 2, which is trained with medical text and offers enterprise grade privacy, security and governess.

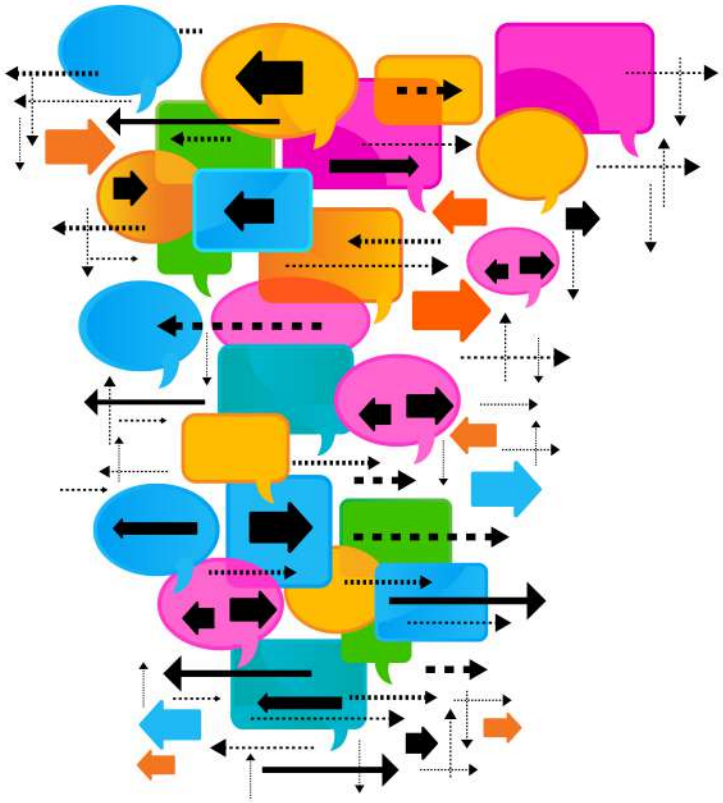
To fully leverage the benefits of LLM, practitioners should take full advantage of its greatest attribute, that being the codeless (or low code) development of applications, and obtain the highest accuracy in response, while avoiding the known pitfall of hallucination, which occurs when the subject of inquiry was not included in the model's training data.

Let's quickly go over the various approaches for implementing LLM. The diagram below illustrates 5 such paths of development.



The last two options in the graph - fine tuning and foundation model training, are the most demanding in resources, and in many cases are simply out of reach for the business owners. We will therefore focus on the first three.

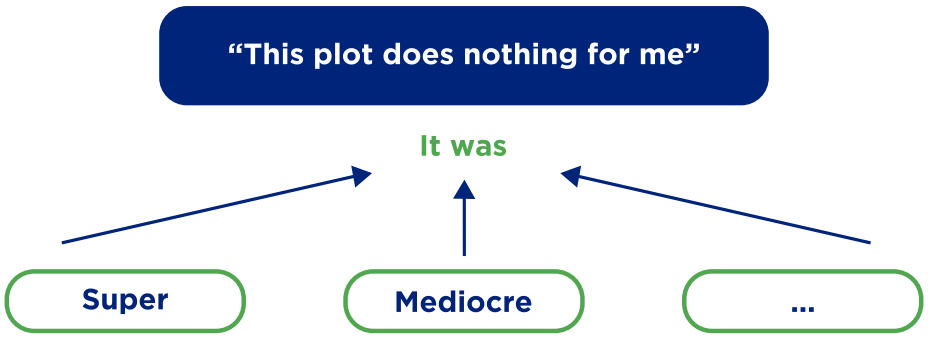
Prompt Engineering



The native interface for an LLM is Question and Answer (Q&A). Therefore, the most direct means of interacting with such a model is simply submitting questions along with an input source. The mechanism of prompting allows the user to set the context, shape the template, refine the response, and even cast the LLM with a desired persona, such as a critic, journalist, or operator, to fine tune the response into the most appropriate form possible. Over time, many prompt engineering techniques have been developed, including few-shot, least to most, chain of thought, self-consistency, and many more.

The chain of thought (CoT) is one of the most interesting approaches in that you are actually walking the LLM through a rather complicated path of reasoning to get to an answer, which is usually some kind of mathematical derivation. An example would be if Tom has 6 tennis balls. He buys 2 new cans with 3 balls in each, then gives 10 balls to his friend Ann. How many balls does Tom have left? The use would then give the LLM an example of how to solve the problem with simple arithmetic operations to derive the answer of 2. This capability highlights the fact that LLM is actually, able to exercise reasoning.

Nevertheless, questions that involve complex reasoning can trip up the LLM. A simple means of countering false responses is by means of consistency prompting, where the user asks the same question a few times, and takes the majority answer as the most reliable one. Prompting can be further polished to implement classification using nothing more than a conversational response involving a template and selective keywords. For example, if one wants to know if a particular movie review is good or bad, a template like the following;



Where the first sentence is the input review text, and the response template follows in the form of "It was _____". With super or mediocre..., representing the binary classes of good and bad. The LLM's performance can vary widely with the choice of template and keywords depending on training data used. However, there are ways to generate the most optimal template and keyword set for the particular instance of LLM to obtain the highest classification accuracy.

Retrieval Augmented Generation

ChatGPT has earned a bad name for answering questions it doesn't know with equal confidence. This phenomenon is known as hallucination, and has become the focus of lots of on-going research to overcome this shortcoming.

The best way to overcome hallucination is to combine the LLM generation capability with information retrieval using a vector database or knowledge graph. The retrieval part leverages techniques like embedding or term frequency - document frequency (TF-IDF) to collect relevant content and feed into LLM, which in term generates the response in accordance with characteristics laid out in the user prompt.

The two methods for integration are either answer first or look up first.

Answer first

The user obtains the first answer from LLM with prompt techniques such as CoT, then double check the truthfulness of the response by querying the attached database.

Look up first

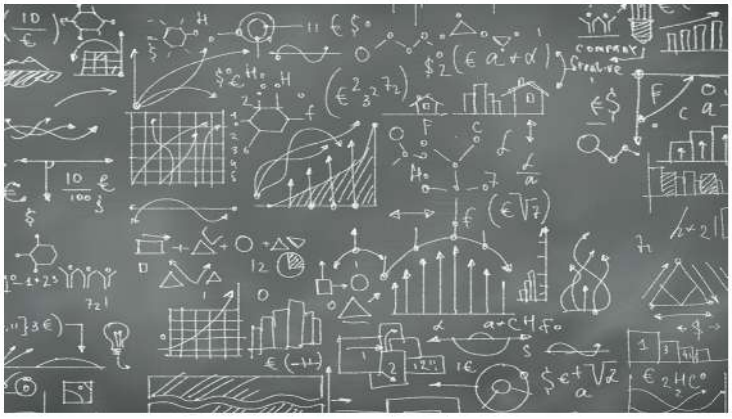
The user obtains facts from the database first, and generates a prompt based on the returned content.

These two methods can be linked into an iterative workflow, so that the answers from LLM are repeatedly verified with database queries, and ends up with a final answer. The user can in fact enrich the database with answers generated from this process if the confidence is sufficiently high.

In a technique known as FLARE (Forward-Looking Active REtrieval augmented generation), it actively scores LLM responses with a confidence level by comparing the response with predicted sentences, and directs low confidence answers to further database retrievals.

Integration with Weak Supervision

Lastly, we can shun potential hallucinations by actively inserting human designs with weak supervision labeling functions (LF). Weak supervision was originally created to circumvent the extreme laborious process of manual labeling training samples. While it has been shown that LLM can already speed up labeling by orders of magnitude, such a method still processes the dataset one sample at a time. Labeling functions, on the other hand, processes all samples in one fell swoop, and also incorporates semi-supervised techniques to leverage unlabeled samples to improve classification accuracy.



Labeling functions creates soft rules that are not 100% correct, and rely on recent advances in semi-supervised learning techniques to boost the overall accuracy to the level of supervised learning. That is a subject for another paper. Here, we simply discuss the merits of using LLM to implement labeling functions to boost the per LF accuracy. For example, it is typical to see LF using a best guess numerical distance to detect a certain pattern, e.g. when the word "not" is within 5 word distance from a keyword, we conclude the keyword is negated. LLM removes the uncertainty of the hardcoded numerical setting, and therefore can detect negation much more reliably.

BigRio is best in industry for delivering the Benefits of LLM to Business Practices

BigRio is a technology consulting firm empowering data to drive innovation and advanced AI. We specialize in cutting-edge Big Data, Machine Learning, LLM and Custom Software strategy, analysis, architecture, and implementation solutions.

Building your own LLM model or looking for a customized solution, there will be many ethical considerations and challenges. While the potential benefits of LLMs in health care are substantial, ethical considerations and challenges must be addressed. Patient privacy, data security, and bias in AI algorithms are critical concerns that need to be carefully managed. Striking the right balance between automation and human intervention is also important to ensure that LLMs are used as tools to augment and support health care professionals rather than replacing them. The BigRio team has an outstanding track record of addressing these challenges and concerns.

Please email us at info@bigr.io to learn more about how we can collaborate and continue to move the needle forward using AI & LLM solutions.