AI in a Nutshell

The world we live in is fast changing due to artificial intelligence (AI). AI has a clear impact on everything from self-driving cars to medical diagnosis. This talk will examine the possibilities, difficulties, and capacities of generative artificial intelligence (AI). We'll look at the underlying mechanics and cutting-edge methods being created to overcome its drawbacks. Get ready to be astounded by this cutting-edge technology's strength and potential.

AI and the Subsets Diagram

Artificial Intelligence (AI)

The wide field that includes the creation of autonomously reasoning, learning, and acting systems, or intelligent agents. This covers a broad spectrum of methods and strategies.

Machine Learning (ML)

A branch of AI that concentrates on giving systems the ability to learn from data without the need for explicit programming. Based on the supplied data, machine learning algorithms find patterns and forecast future events.

Deep Learning (DL)

A branch of machine learning that extracts higher-level features from data using multi-layered artificial neural networks. Large datasets and intricate patterns are ideal for deep learning applications.

Introducing Generative AI (GenAI)

1 What is GenAI?

Artificial intelligence that can produce original text, graphics, music, and video is known as generative AI. GenAI concentrates on generation, as opposed to standard AI systems, which concentrate on analysis and prediction..

2 Foundation Models

These are large-scale artificial intelligence models that have been trained on enormous datasets and can accomplish a variety of tasks with little more training. They serve as the basis for numerous GenAI applications.

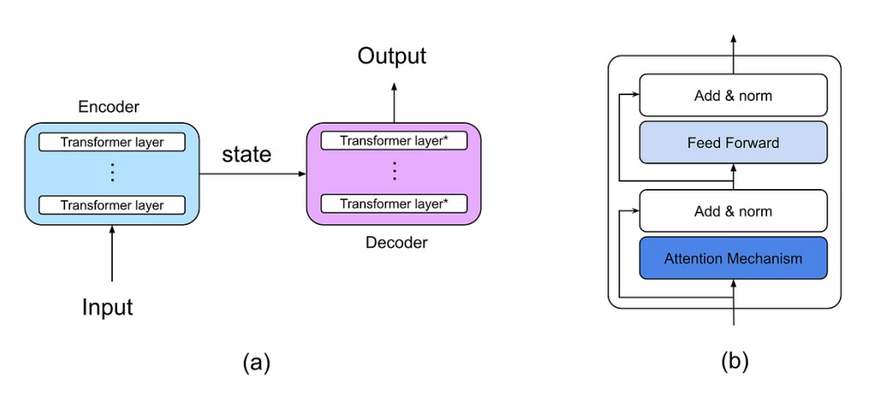
3 Large Language Models (LLMs)

A particular kind of foundation model that can comprehend and produce text that is human-like after being trained on enormous volumes of text data. Many text-related GenAI applications are based on LLMs.

4 Open vs. Closed Models

Open models are available to the whole public and welcome community input and enhancements. Closed models are exclusive to their creators and are proprietary.

Large Language Models (LLMs) and Foundation Models



1

Foundation Models

These are strong, already-trained AI models that form the foundation of many different applications. Having been taught on large datasets, they require little additional training to accomplish a wide range of tasks.

2

Language Models

A particular kind of foundation model that was trained on text data in order to comprehend and produce text that is human-like. These models provide coherent and contextually relevant output by learning patterns in language.

3

Large Language Models (LLMs)

These language models are extraordinarily massive, containing billions or even trillions of parameters. Their size enables them to produce excellent text, translations, and more by capturing intricate linguistic patterns.

How LLMs Operate: Information to Creation

1

Ingestion of Data

Massive text and code datasets are used to train LLMs; these datasets are then divided into smaller chunks known as tokens.

2

Using Chunking and Tokenization

Tokens are processed in groups called chunks. After that, these pieces are transformed into numerical representations known as embeddings.

3

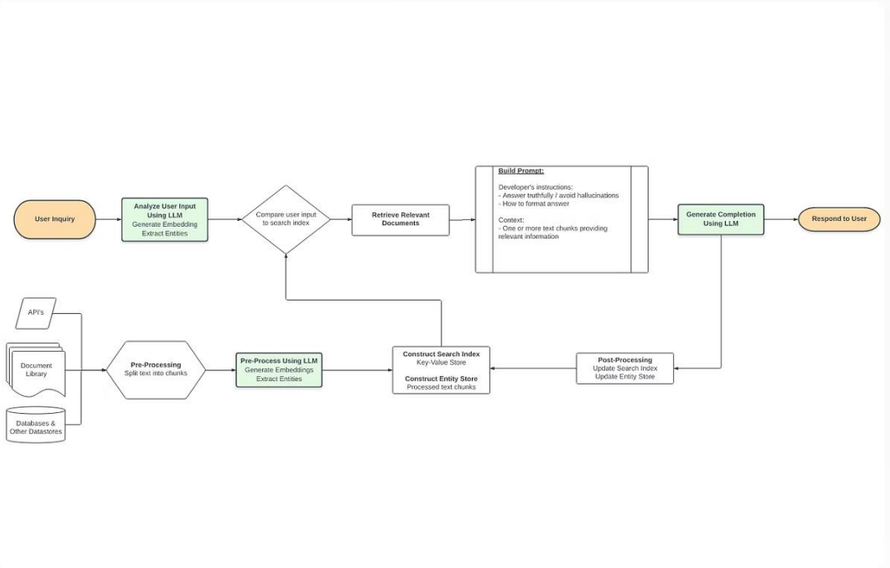
Storage and Embedding

Vector databases (VDBs) are used to store embeddings for effective retrieval. The model creates embeddings for prompts when they are presented.

4

Generation and Similarity of Cosines

By utilizing cosine similarity to identify the most comparable embeddings in the VDB, the model predicts the subsequent token in the sequence and generates the response.



The Issue of Hallucinations in LLMs

A hallucination is what?

In LLMs, creating information that is illogical or factually inaccurate is referred to as hallucination. This might be anything from small typos to wholly made up claims.

Reasons for Delusions

A number of reasons, such as biases in the training data, limits in the model's contextual awareness, and the intrinsic statistical structure of the generation process, might lead to hallucinations.

Effects of Delusion

Misinformation, erroneous forecasts, and a decline in confidence in the LLM can result from hallucinations. It is imperative that this crucial issue be resolved before this technology is widely adopted.

How to Get Rid of Hallucinations: PE, FT, RAG

Prompt Engineering (PE)

creating prompts with care to direct the LLM toward correct answers. This entails giving precise directions, pertinent background, and limitations.

Fine-tuning (FT)

To enhance the LLM's performance on certain tasks and lessen hallucinations in those domains, more training on a more specialized dataset is required.

Retrieval Augmented Generation (RAG)

integrating LLMs with outside knowledge sources to provide the generated text with accurate information as a foundation. By doing this, it is ensured that the LLM's answers match the information supplied.

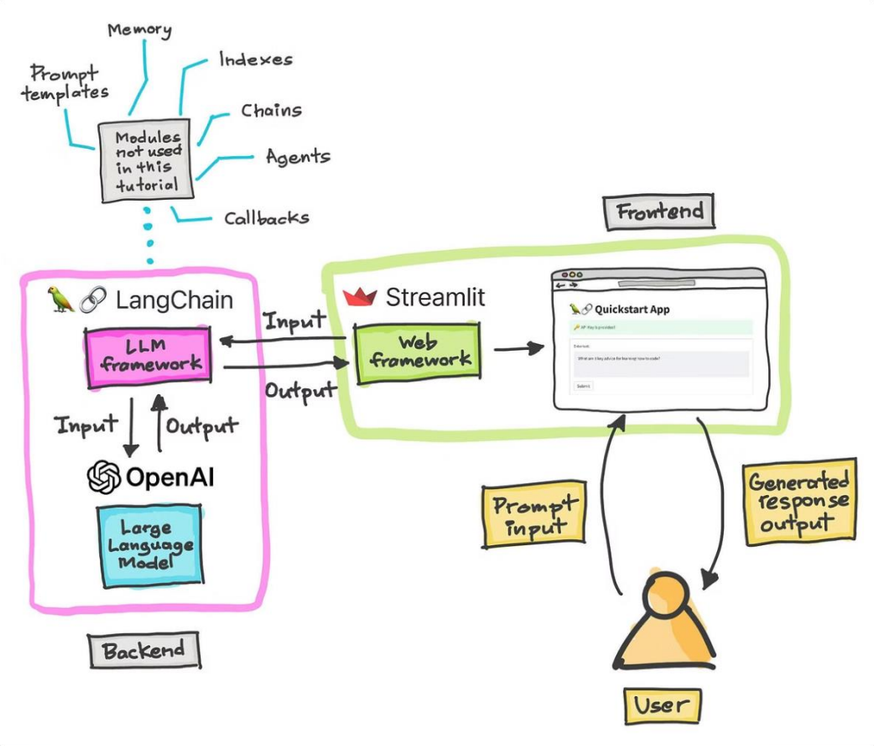
Tooling for GenAI

Sidebar 2 - Software Development Productivity Drivers

Driver	Rationale	Impact on productivity: Low	Average	High
Custom software	Buy or tailor-made software where GenAI can propose innovative functionalities, generate code, and assist with design	Mostly COTS[1]	Balanced mix	Mostly in house
Sourcing model	Strategic approach to acquire IT resources, services, and solutions; in- house model allows more flexibility to integrate GenAI	Outsourced	Time and material	Internal developers
Hosting	Reliable and scalable infrastructure is key to create an effective environment for running GenAI models	On premise	Hybrid	Public cloud
DevOps	DevOps practices enable continuous integration and deployment to ensure GenAI model is always up to date, automatically tested, and deployed	No DevOps	Partially DevOps	Full DevOps
Development methodology	Allows continuous improvements with regular feedbacks and transparency to refine models and adjust enabling faster time to market	V cycle	Agile	Fully agile
Coding language	Modern coding languages have extensive libraries and frameworks supporting already AI and machine learning tasks	Niche legacy (Cobol)	Classic (C++, Java)	Modern (JS, Python)

Source: BCG analysis.
[1]COTS = commercial off-the-shelf software packages.

Tool	Description
LangChain	a framework for creating LLM-based programs that offers memory management, chain sequencing, and prompt management capabilities.
LlamaIndex	An indexing architecture that makes it possible to link LLMs to your own data so that you can respond with accuracy and personalization.
Vertex AI	A variety of resources and services, including LLMs, are available for creating, honing, and implementing AI models on Google Cloud's machine learning platform.



LangChain

1 Sections

For varied functionalities, such as chains, memory, indexes, and prompt templates, LangChain offers a variety of modules.

2 Chains

Chains facilitate the development of intricate workflows by merging several LLMs or other elements to accomplish a specific goal.

3 Indexes

By facilitating access to and processing of external data sources, indexes help LLMs produce writing that is more accurate and pertinent.

4 Memory

Memory components let LLMs recall past conversations and keep a consistent identity by preserving context across many interactions.

Conclusion and Key Takeaways

Artificial intelligence has advanced significantly with generative AI, especially LLMs. Although there are difficulties such as hallucinations, methods such as PE, FT, and RAG are being developed to deal with these problems. Building apps utilizing LLMs is made easier by tools like as LangChain. AI has a bright future ahead of it, with the potential to completely transform a wide range of elements of our lives. In order to minimize the hazards associated with this revolutionary technology and realize its full potential, ongoing research and development will be essential.