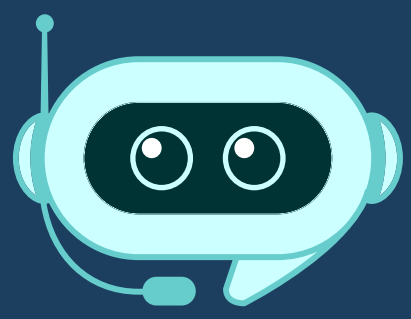
datasciencedojo

data science for everyone

Master These 9 Areas to Excel



as An LLM Expert

https://datasciencedojo.com/

datasciencedojo
data science for everyone

Generative AI

Generative AI is a category of AI algorithms that generate new content based on the data they have been trained on. Generative language models learn about patterns in language through training data.

There are two major types of Generative AI:

- Image-based models: These models employ methods such as diffusion to create new images. e.g. MidJourney
- Text-based models: These models have the ability to generate human-like textual content. e.g. ChatGPT

Embeddings

Embeddings are simply numerical representations of words or concepts that capture their semantic relationships in a multi-dimensional space, enabling algorithms to understand and compare textual information.

In the case of GPT-3, each word or token is represented as a high-dimensional vector, strategically positioned in the vector space to capture word similarities. This arrangement aids the model in:

- Understanding word relationships
- Enabling the generation of coherent and contextually relevant text.

datasciencedojo
data science for everyone

Vector Databases

Vector databases store pre-computed vectors from a wide range of text, enabling efficient retrieval and comparison. When a query is made, the LLM can use these vectors to find relevant content quickly. They are used in a variety of AI applications, such as image search, natural language processing, and recommender systems. The pipeline for a vector database includes:

- Indexing: The indexing step is responsible for mapping the vectors to a data structure that will enable faster searching
- Querying: The querying step is responsible for finding the nearest neighbors to a query vector
- Post-processing: The post-processing step is optional. It can be used to improve the results of the query by re-ranking the nearest neighbors

datasciencedojo
data science for everyone

Prompt Engineering

Well-crafted prompts guide the LLM's output. They provide context, instructions, or queries to influence the generated response's relevance and accuracy. Here are some basic prompt engineering techniques you must know:

- Be clear and specific
- Provide context
- Experiment with different prompts
- Use relevant keywords
- Refine the prompt

Semantic Search

Semantic search is a breakthrough in information retrieval, it allows the system to do so. It grasps user intent beyond mere **keywords**. This happens through vector databases where data is encoded into vectors. When you query, the system seeks answers with high cosine similarity to your input vector.

This approach ensures:

- Robust information retrieval
- Enhancing search precision

datasciencedojo
data science for everyone

Orchestration Frameworks

Orchestration Frameworks are responsible for allowing an LLM to answer queries based on the custom data that you provide to it. These frameworks handle the sequencing of inputs, responses, and actions, guaranteeing a smooth and contextually appropriate user experience. There are a number of orchestration frameworks available, here are the two most popular ones:

- LangChain is an open-source orchestration framework that is designed to be easy to use and scalable.
- Llama Index is another open-source orchestration framework that is designed for managing LLMs output responses.

datasciencedojo
data science for everyone

LangChain Agents

LangChain agents use an LLM to decide what actions to take and the order to take them, making future decisions by iteratively observing the outcome of prior actions. Agents can be chained together and they can connect the LLM to external knowledge sources or tools for computation. Here are 3 different types of agents:

- OpenAPI agent: It helps consume arbitrary APIs that conform to the OpenAPI/Swagger specification.
- CSV agent: It is an agent built on top of the Pandas DataFrame agent capable of querying structured data and question-answering over CSVs. It loads data from CSV files and can perform basic querying operations like selecting and filtering columns, sorting data, and querying based on a single condition.
- Pandas dataframe agent: It is an agent built on top of the Python agent capable of question-answering over Pandas dataframes, processing large datasets by loading data from Pandas dataframes, and performing advanced querying operations.

Fine-tuning Large Language Models

Pre-trained LLMs offer impressive capabilities like text generation, summarization, and coding. However, they aren't universally fitting for every scenario especially when the information required is relevant to a specific domain. Fine-tuning helps in specialized tasks. It involves enhancing the foundational model with specialized data in domains like medicine, finance, law, etc. Here are different types of fine-tuning:

- Instruction fine-tuning
- Unsupervised fine-tuning
- RLHF fine-tuning

datasciencedojo
data science for everyone

LLMOps

LLMOps encompasses the practices, techniques, and tools used for the operational management of large language models in production environments. It helps with:

- Scaling
- Maintenance
- Monitoring
- Observability
- Guardrails

LLMOps ensures the application's reliability, stability, and optimal performance over time.

datasciencedojo
data science for everyone

Want to Build a Custom LLM Application?

Learn More

Check out our in-person Large Language Models bootcamp.

Seattle

January 29 - February 2, 2024

Online

Coming Soon

https://datasciencedojo.com/