datasciencedojo
data science for everyone

Building LLM Applications

A 50-hour roadmap

A comprehensive roadmap to building large language model applications in ~50 hours

Data Science for Everyone

https://datasciencedojo.com

Introduction to Generative AI

Quick overview of generative AI, LLMs, and foundation models. Learn more about how transformers and attention mechanism works behind the text and image-based models:

- Types of generative AI models
  - Text-based models
  - Image-based models
- Foundation models & LLMs
  - Encoder decoder
  - Attention mechanism
  - Transformers model and BERT model
- Intro to Image Generation
  - Image captioning models
  - Diffusion models
- Generative AI applications
  - ChatGPT & Bard
  - DALL-E & Midjourney

Emerging Architectures

Understand the common use cases of large language models and the fundamental building blocks of such applications. Learners will be introduced to the following topics at a very high level without going into the technical details:

- Large language models and foundation models
- Vector databases, embeddings, and LLM cache
- Prompts and prompt engineering
- Context window and token limits
- Embeddings and vector databases
- Build custom LLM applications by:
  - Training a new model from scratch
  - Fine-tuning foundation LLMs
  - In-context learning
- Canonical architecture for an end-to-end LLM application

Embeddings

In this module, we will be reviewing how embeddings have evolved from the simplest one-hot encoding approach to more recent semantic embeddings approaches. The module will go over the following topics:

- Review of classical techniques
  - Review of binary/one-hot, count-based, and TF-IDF techniques for vectorization
  - Capturing local context with n-grams and challenges
- Semantic encoding techniques
  - Overview of Word2Vec and dense word embeddings
  - Application of Word2Vec in text analytics and NLP tasks
- Hands-on exercise
  - Creating a TF-IDF and semantic embeddings on a document corpus

Attention Mechanism and Transformers

Dive into the world of large language models, discovering the potent mix of text embeddings, attention mechanisms, and the game-changing transformer model architecture. This module consists of:

- Text embeddings
  - Word and sentence embeddings
  - Multilingual sentence embeddings
- Text similarity measures
  - Dot product, cosine similarity, inner product
- Hands-on exercise
  - Calculating similarity between sentences using cosine similarity and dot product
- Attention mechanism and transformer models
  - Neural machine translation (NMT) and sequence-to-sequence models
  - Attention mechanism components
  - Self-attention and multi-head attention
  - Transformer networks: Tokenization, embedding, positional encoding, and transformers block
- Hands-on exercise
  - Understanding attention mechanisms: Self-attention for contextual word analysis

Vector Databases

Learn about efficient vector storage and retrieval with vector databases, indexing techniques, retrieval methods, and hands-on exercises:

- Overview
  - Rationale for vector databases
  - Importance of vector databases in LLMs
  - Popular vector databases
- Indexing techniques
  - Product quantization (PQ), Locality sensitive hashing (LSH), and Hierarchical navigable small world (HNSW)
- Retrieval techniques
  - Cosine similarity
  - Nearest neighbor search
- Hands-on exercise
  - Creating a vector store using HNSW
  - Creating, storing, and retrieving embeddings using cosine similarity and nearest neighbors

Semantic Search

Understand how semantic search overcomes the fundamental limitation in lexical search i.e. lack of semantics. Learn how to use embeddings and similarity in order to build a semantic search model:

- Understanding and implementing semantic search
  - Introduction and importance of semantic search
  - Distinguishing semantic search from the lexical search
  - Semantic search using text embeddings
- Exploring advanced concepts and techniques in semantic search
  - Multilingual search
  - Limitations of embeddings and similarity in semantic search
  - Improving semantic search beyond embeddings and similarity
- Hands-on exercise
  - Building a simple semantic search engine with multilingual capability

Prompt Engineering

Unleash your creativity and efficiency with prompt engineering. Seamlessly prompt models, control outputs, and generate captivating content across various domains and tasks. This module includes:

- Prompt design and engineering
    - Prompting by instruction
    - Prompting by example
- Controlling the model output
    - When to stop
    - Being creative vs. predictable
    - Saving and sharing your prompts
- Use case Ideation
    - Utilizing goal, task, and domain for perfect prompt
- Example use cases
    - Summarizing (summarizing a technical report)
    - Inferring (sentiment classification, topic extraction)
    - Transforming text (translation, spelling, and grammar correction)
    - Expanding (automatically writing emails)
    - Generating a product pitch
    - Creating a business model canvas
    - Simplifying technical concepts
    - Composing an email

Fine-Tuning Foundation Models

Discover the ins and outs of fine-tuning foundation language models (LLMs) through theory discussions, exploring rationale, limitations, and parameter efficient fine-tuning (PEFT):

- Fine-tuning foundation LLMs
  - Rationale for fine-tuning
  - Limitations of fine-tuning
  - Parameter efficient fine-tuning
- Hands-on exercise
  - Fine-tuning and deploying the OpenAI GPT model on Azure

Orchestration Frameworks

Explore the necessity of orchestration frameworks, tackling issues like foundation model retraining, token limits, data source connectivity, and boilerplate code. Discover popular frameworks, their creators, and open-source availability:

- Why are Orchestration Frameworks (OF) needed?
  - Eliminate the need for foundation model retraining
  - Overcoming token limits
  - Connecters for data sources

LangChain

Build LLM apps using LangChain. Learn about LangChain's key components such as models, prompts, parsers, memory, chains, and QnA. Get hands-on evaluation experience:

- Introduction to LangChain
  - Schema, models, and prompts
  - Memory and chains
- Loading, transforming, indexing, and retrieving data
  - Document loader
  - Text splitters
  - Retrievers
- LangChain use cases
  - Summarization: Summarizing long documents
  - QnA using documents as context
  - Extraction: Getting structured data from unstructured text
  - Evaluation: Evaluating outputs generated from LLM models
  - Querying tabular data without using any extra code
- Hands-on exercise
  - Using LangChain loader, splitter, and retrievals on a pdf document

Autonomous Agents

Use LLMs to make decisions about what to do next. Enable these decisions with tools. We'll learn what they are, how they work, and how to use them within the LangChain library to superpower our LLMs. In this module, we'll talk about:

- Agents and tools
- Agent types
  - Conversational agents
  - OpenAI functions agents
  - ReAct agents
  - Plan and execute agents
- Hands-on exercise: Create and execute some of the following agents
  - Excel agent
  - JSON agent
  - Python Pandas agent
  - Document comparison agent
  - Power BI agent

Bias, Fairness and Explainablity

Bias can creep in at any stage of the lifecycle of a model. While large language models offer tremendous business value, humans are involved in all stages of the lifecycle of an LLM from acquisition of data to interpretation of insights. In this module, we will learn about the following:

- Ethics, bias, fairness
  - Sources of bias in acquisition/annotation of training data, model building
  - Precautions against safeguarding the model from bias
- Review some of the regulations/legislation
- Principles of responsible AI
  - Fairness and eliminating bias
  - Reliability and safety
  - Privacy and data protection
  - Transparency and explainability
  - Accountability and governance
  - Inclusivity and accessibility
- Review some of the tools available to assess the following in a large language model application
  - Correctness and security
  - Bias, fairness, and explainability of the model

Recommended Projects

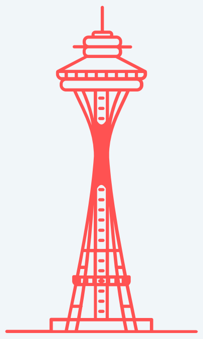
- Virtual assistant: A dynamic customer service agent designed for the car manufacturing industry.
- Content generation (Marketing co-pilot): Enhancing your marketing strategies with an intelligent co-pilot.
- Conversational agent (Legal and compliance assistant): Assisting with legal and compliance matters through interactive conversations.
- QnA (IRS tax bot): An intelligent bot designed to answer your questions about IRS tax-related topics.
- Content personalizer: Tailoring content specifically to your preferences and needs.
- YouTube virtual assistant: Engage in interactive conversations with your favorite YouTube channels and playlists.

datasciencedojo
data science for everyone

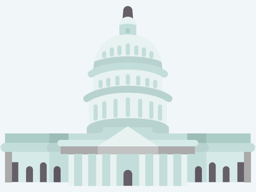
Learn to Build LLM Applications

Join this 5-day | 40-hour bootcamp to get started with building large language model applications on your enterprise data

Learn More



Seattle
September 18-22, 2023



Washington, D.C.
October 16-20, 2023

Austin
November 6-10, 2023

New York
December 4-8, 2023

Singapore
January, 2024