Chat models: LLMs exposed via a chat API that process sequences of messages as input and output a me
Messages: The unit of communication in chat models, used to represent model input and output.
Chat history: A conversation represented as a sequence of messages, alternating between user messages
Tools: A function with an associated schema defining the function's name, description, and the arguments
Tool calling: A type of chat model API that accepts tool schemas, along with messages, as input and retur
Structured output: A technique to make a chat model respond in a structured format, such as JSON that m
Memory: Information about a conversation that is persisted so that it can be used in future conversations.
Multimodality: The ability to work with data that comes in different forms, such as text, audio, images, and
Runnable interface: The base abstraction that many LangChain components and the LangChain Expressio
Streaming: LangChain streaming APIs for surfacing results as they are generated.
LangChain Expression Language (LCEL): A syntax for orchestrating LangChain components. Most useful
Document loaders: Load a source as a list of documents.
Retrieval: Information retrieval systems can retrieve structured or unstructured data from a datasource in r
Text splitters: Split long text into smaller chunks that can be individually indexed to enable granular retrieva
Embedding models: Models that represent data such as text or images in a vector space.
Vector stores: Storage of and efficient search over vectors and associated metadata.
Retriever: A component that returns relevant documents from a knowledge base in response to a query.
Retrieval Augmented Generation (RAG): A technique that enhances language models by combining them
Agents: Use a language model to choose a sequence of actions to take. Agents can interact with external
Prompt templates: Component for factoring out the static parts of a model "prompt" (usually a sequence of
Output parsers: Responsible for taking the output of a model and transforming it into a more suitable forma
Few-shot prompting: A technique for improving model performance by providing a few examples of the tas
Example selectors: Used to select the most relevant examples from a dataset based on a given input. Exa
Async programming: The basics that one should know to use LangChain in an asynchronous context.
Callbacks: Callbacks enable the execution of custom auxiliary code in built-in components. Callbacks are u
Tracing: The process of recording the steps that an application takes to go from input to output. Tracing is
Evaluation: The process of assessing the performance and effectiveness of AI applications. This involves t
Testing: The process of verifying that a component of an integration or application works as expected. Tes