Language models

LangChain provides interfaces and integrations for two types of models:

- LLMs: Models that take a text string as input and return a text string
- Chat models: Models that are backed by a language model but take a list of Chat Messages as input and return a Chat Message

LLMs vs Chat Models

LLMs and Chat Models are subtly but importantly different. LLMs in LangChain refer to pure text completion models. The APIs they wrap take a string prompt as input and output a string completion. OpenAI's GPT-3 is implemented as an LLM. Chat models are often backed by LLMs but tuned specifically for having conversations. And, crucially, their provider APIs expose a different interface than pure text completion models. Instead of a single string, they take a list of chat messages as input. Usually these messages are labeled with the speaker (usually one of "System", "AI", and "Human"). And they return a ("AI") chat message as output. GPT-4 and Anthropic's Claude are both implemented as Chat Models.

To make it possible to swap LLMs and Chat Models, both implement the Base Language Model interface. This exposes common methods "predict", which takes a string and returns a string, and "predict messages", which takes messages and returns a message. If you are using a specific model it's recommended you use the methods specific to that model class (i.e., "predict" for LLMs and "predict messages" for Chat Models), but if you're creating an application that should work with different types of models the shared interface can be helpful.