Language Models

Understanding AI Models for Text and Conversational Tasks

Overview

- **Language Models:** AI systems designed to process, generate, and understand natural language text.
- **LLMs (General-purpose models):** Take plain text as input and produce plain text outputs. Older models, mainly used for raw text generation.
- **Chat Models:** Specialized for conversational tasks. Take a sequence of messages as input and return structured chat messages. These are newer and more widely used in chat applications.

1. Comparison: LLMs vs Chat Models

Feature	**LLMs (Base Models)**	**Chat Models (Instruction-Tuned)**
Purpose	Free-form text generation	Optimized for multi-turn conversations
Training Data	General text corpora (books, articles)	Fine-tuned on chat datasets (dialogues, user-assistant conversations)
Memory & Context	No built-in memory	Supports structured conversation history
Role Awareness	No understanding of "user" and "assistant" roles	Understands "system", "user", and "assistant" roles
Example Models	GPT-3, Llama-2-7B, Mistral-7B, OPT-1.3B	GPT-4, GPT-3.5-turbo, Llama-2-Chat, Mistral-Instruct, Claude
Use Cases	Text generation, summarization, translation, creative writing, code generation	Conversational AI, chatbots, virtual assistants, customer support, AI tutors

LangChain Note

- In **LangChain**, LLMs inherit from `BaseLLM`, providing standard methods for text-based prompts.

- Chat Models inherit from `BaseChatModel`, providing structured interfaces for multi-turn conversation messages and role-aware interactions.

Key Takeaways

- Use **LLMs** for raw text tasks such as summarization, translation, or code generation.
- Use **Chat Models** for interactive, multi-turn conversations and applications that require role awareness and context.
- Modern applications typically rely more on chat models due to their enhanced conversational capabilities.