COURSE GLOSSARY

Prompt Engineering with the OpenAI API

API key: A secret token provided by an API service that authenticates and authorizes a user or application to send requests to the service

Assistant message: A message role representing the model's responses to user prompts or previous assistant content in a conversation

Chain-of-thought prompting: Instructing the model to generate its intermediate reasoning steps or "thoughts" before producing a final answer to increase transparency and improve performance on complex problems

Chat completions endpoint: An API endpoint that accepts a sequence of messages (with roles) and returns the model's conversational response, commonly used to build chatbots

Delimited prompt: A prompt structure that uses clear delimiters (e.g., backticks, brackets) to separate instructions from input data so the model can reliably locate and operate on the input

Entity extraction: The process of identifying and extracting structured entities (like names, dates, product names, or locations) from unstructured text for downstream use

External context (in prompts): Additional information provided to the model—such as company facts, prior conversations, or documents—so it can answer questions beyond its pretrained knowledge

Large language model (LLM): A neural network trained on large amounts of text data that can generate or analyze human-like text given prompts

Max_tokens: A parameter that limits the number of output tokens the model can generate, effectively bounding response length and potentially truncating long outputs

Multi-step prompting: Breaking a complex task into a sequence of explicit sub-steps within the prompt so the model completes each step in order to improve correctness and coherence

One-shot prompting: Providing exactly one example in the prompt to demonstrate the desired input-output format or style for the model to imitate

Prompt engineering: The practice of designing and refining the instructions (prompts) given to a language model to elicit accurate, useful, and appropriately formatted responses

Role-playing prompt: A prompt that instructs the model to adopt a specific persona or professional role (e.g., product manager, sales engineer) to shape tone, focus, and domain knowledge in responses

Self-consistency prompting: A technique that samples multiple chain-of-thought style responses and aggregates them (e.g., by majority vote) to improve robustness and reduce reasoning errors

Structured output: A prompt design requirement that tells the model to produce output in a specific format (tables, lists, JSON, headings) to make downstream parsing or consumption easier

System message: A message role in chat-based APIs used to give global instructions that shape the model's behavior throughout a conversation

Temperature: A sampling parameter (commonly 0–2) that controls randomness in model outputs, where lower values make responses more deterministic and higher values increase variability

Text summarization: The task of condensing a longer piece of text into a shorter version that preserves the main ideas and essential information

User message: A message role representing the end user's input or prompt that asks the model to perform a task or provide information

Zero-shot prompting: Asking a model to perform a task without providing any examples, relying solely on the instruction and the model's pre-existing knowledge