Concept | Large language models and the LLM Mesh

Large Language Models (LLMs) are a type of Generative AI that specializes in generating coherent responses to input text, known as prompts.

With an extensive understanding of language patterns and nuances, LLMs can be used for a wide range of natural language processing tasks, such as classification, summarization, translation, question answering, entity extraction, spell check, and more.

LLMs have transformed how many industries use AI, with seemingly endless use cases from writing and coding assistants to specialized chatbots and beyond.

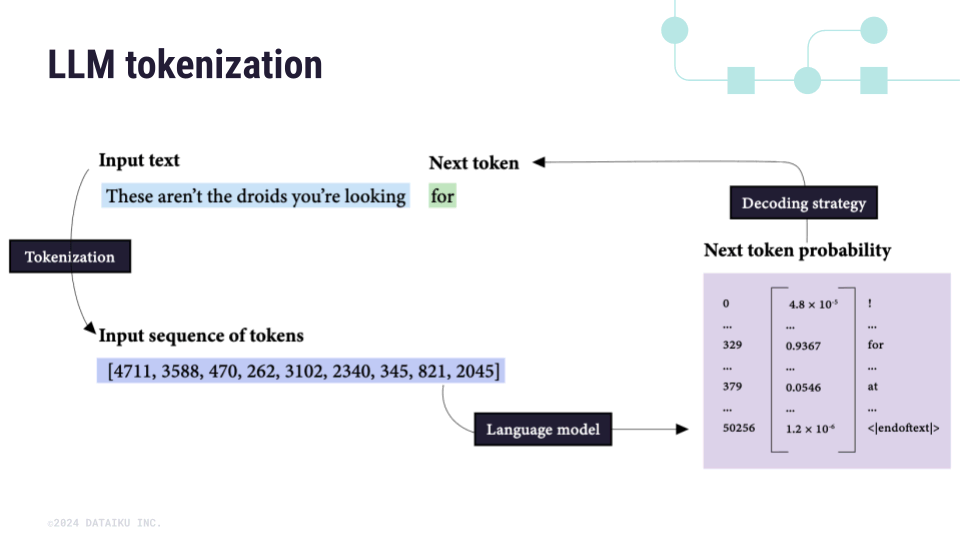
In Dataiku, you can connect to commercially available LLMs, host your own locally, and even fine-tune specialized LLMs. You can work with LLMs through visual recipes or code, and coordinate connections to multiple models through Dataiku’s LLM Mesh architecture.

Let’s take a high-level look at how LLMs work and how you can leverage them in Dataiku!

# LLM basics

LLMs were trained with massive amounts of textual data such as books, articles, and websites. Because training an LLM is a laborious, time-consuming, and costly task, many companies use commercially available pretrained models such as GPT from OpenAI, Claude from Anthropic, or Llama from Meta.

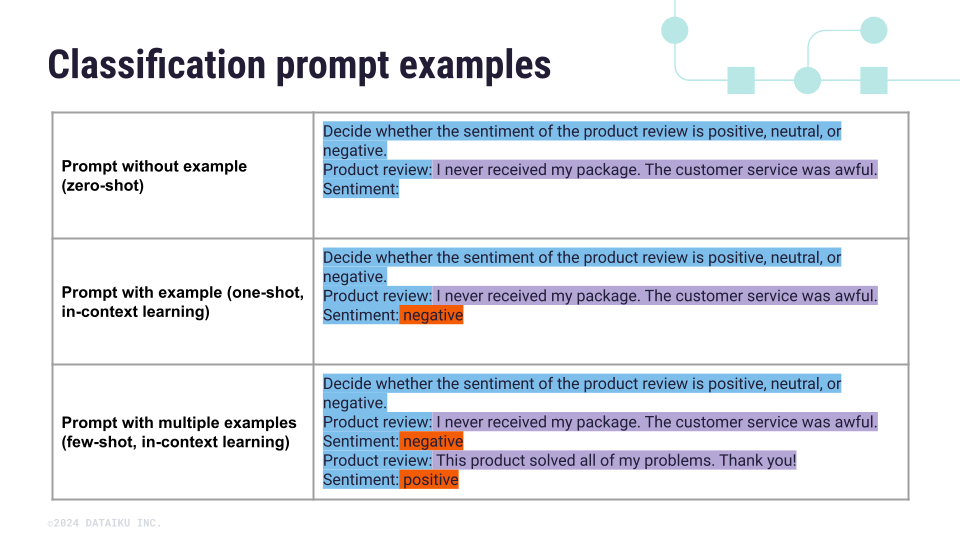
At the most basic level, LLMs take an input text, known as a prompt, and break the input into a sequence of numerical tokens, a process known as tokenization. The model then uses a probability distribution to predict the next token. It continues calculating the probabilities of each token, one by one, in an output sequence.



The key to accomplishing natural language tasks with LLMs is to craft a prompt in a way that the continuation of this prompt yields the desired answer.

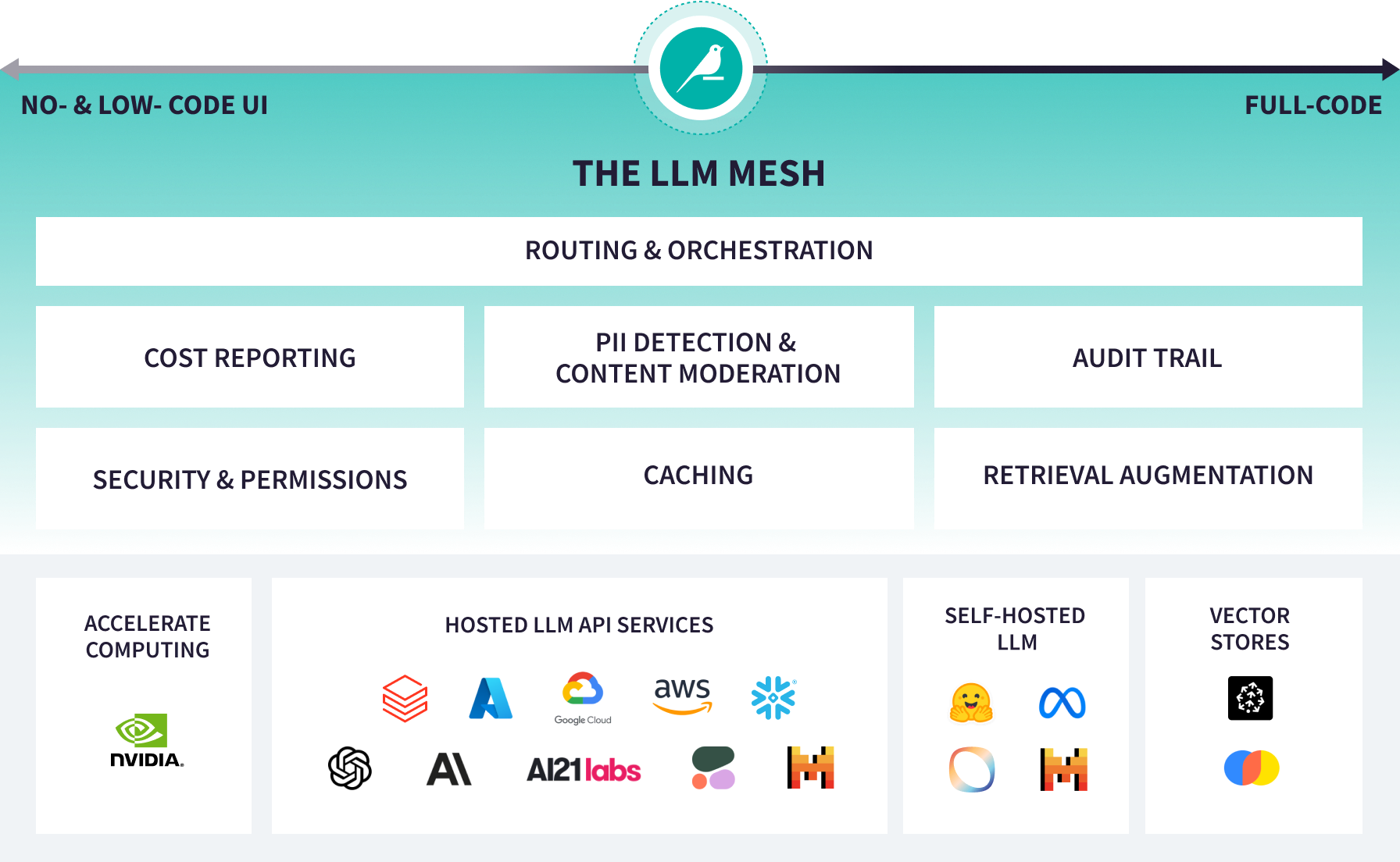
For example, suppose you want to classify a product review as positive, negative, or neutral. A simple prompt would be to specify the task and include the text of the product review. However, you can also include one or more examples of the desired output, showing through the prompt how you’d like the output to be structured.

These strategies, known as zero-shot prompting (without output examples) and one-shot prompting (with one output example), are only two of many ways to approach prompt engineering with LLMs.



# LLM Mesh

In Dataiku, connections to LLM providers are orchestrated through the LLM Mesh. The LLM Mesh enables organizations to build enterprise-grade applications while addressing concerns related to cost management, compliance, and technological dependencies. It also enables choice and flexibility among the growing number of models and providers.



The LLM Mesh allows your Dataiku administrators to:

* Connect to the AI services of your choice, including LLMs and vector databases, using an [LLM connection](https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-llm-connections.html). You can switch between models to test performance and cost without recreating applications.
* Centrally [manage access keys](https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-genai-guardrails.html#llm-access-key-management) to different services to ensure that only approved services and datasets are used.
* Detect and redact [Personally Identifiable Information (PII)](https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-genai-guardrails.html#llm-detect-pii).
* Detect and control the [toxicity](https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-genai-guardrails.html#llm-detect-toxic-language) of responses.
* Control costs by [caching responses](https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-llm-connections.html#llm-cache) to common queries and monitoring cost reports.
* Leverage built-in [retrieval augmentation](https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-rag.html) to enrich queries and responses with internal documentation.

# What’s next?

Learn more about connecting to LLMs, the types of connections, and auditing connections in the [Concept | LLM connections](https://knowledge.dataiku.com/latest/ml-analytics/gen-ai/concept-llm-connections.html).