Large language models (LLMs) on Databricks

October 31, 2023

Databricks makes it simple to access and build off of publicly available large language models.

Databricks Runtime for Machine Learning includes libraries like Hugging Face Transformers and LangChain that allow you to integrate existing pre-trained models or other open-source libraries into your workflow. From here, you can leverage Databricks platform capabilities to fine-tune LLMs using your own data for better domain performance.

See the Hello Dolly blog for an example of an open-source LLM model recreated on Databricks.

In addition, Databricks offers built-in functionality for SQL users to access and experiment with LLMs like Azure OpenAI and OpenAI using AI functions.

In this article:

- Hugging Face Transformers
- LangChain
- AI functions

Hugging Face Transformers

With Hugging Face Transformers on Databricks you can scale out your natural language processing (NLP) batch applications a
fine-tune models for large-language model applications.

The Hugging Face `transformers` library comes preinstalled on Databricks Runtime 10.4 LTS ML and above. Many of the popular NLP models work best on GPU hardware, so you might get the best performance using recent GPU hardware unless you use a model specifically optimized for use on CPUs.

- What are Hugging Face Transformers?
- Fine-tune Hugging Face models for a single GPU
- Model inference using Hugging Face Transformers for natural language processing (NLP)

LangChain

LangChain is available as an experimental MLflow flavor which allows LangChain customers to leverage the robust tools and experiment tracking capabilities of MLflow directly from the Databricks environment.

LangChain is a software framework designed to help create applications that utilize large language models (LLMs) and combine them with external data to bring more training context for your LLMs.

Databricks Runtime ML includes `langchain` in Databricks Runtime 13.1 ML and above.

Learn about Databricks specific LangChain integrations.

In addition, Databricks provides AI functions that SQL data analysts can use to access LLM models, including from OpenAI, directly within their data pipelines and workflows. See AI Functions on Databricks.

AI functions

> **Preview**
>
> This feature is in Public Preview.

AI functions are built-in SQL functions that allow SQL users to:

- Access public large language models like those in Azure OpenAI and Open AI and experiment with them on your company's data.

- Query models hosted by Databricks model serving endpoints from SQL queries.

These functions are only available in public preview on Databricks SQL Pro or Serverless.

- ai_generate_text()
  - Analyze customer reviews with ai_generate_text() and OpenAI.
- ai_query()
