Large Language Model (LLM) POC: Knowledge Base Q&A

Databricks Professional Services

Summary

Large language models (LLMs) are the backbone of many natural language processing (NLP) applications, such as ChatGPT and Google Translate. Most out-of-the-box LLMs are general-purpose models trained on publicly available text. However, many business problems need a specialized language model, augmented and/or trained on domain-specific data sets, to deliver business value. Imagine answering questions based on your organization’s knowledge base, integrated with your Databricks Lakehouse.

Databricks can help you by:

- Advising on how to build your own data set to get the most out of a large language model, such as LLama 2 or MPT
- Building a knowledge base Q&A model prototype:
  - Cleaning data and creating a vector database for retrieval
  - Identifying and leveraging the best-suited open source LLMs
  - Fine-tuning a model with your data for custom embeddings (if necessary)
  - Basic model evaluation
  - Any in-depth model evaluation requires manual inspection of results and collaboration with domain experts
  - Enabling internal teams
- Setting up MLflow AI Gateway and understanding next steps to deploy your Q&A model

Key Outcome

- A data strategy and reference implementation for a knowledge base Q&A model jointly determined by customer and Databricks

Key benefits

- Increase developer productivity
- Reduce time to market
- Improve security from keeping your data and model in-house
- Reduce total cost of ownership (TCO)

Out of scope

- Configuration and integration of non-Databricks products
- Data cleansing associated with building broader data lake
- Any efforts to improve an existing benchmark model have no guarantee that the new model will outperform, especially if the benchmark model currently uses OpenAI
- Any necessary PII removal requires additional implementation time
- Integration of the Q&A bot with front-end application
- Implementing CI/CD and automated deployment infrastructure setup
Prerequisites

- Data must exist in PDFs or in a tabular format, ideally with these columns: title, content and web links (unless mutually agreed). Web scraping is out of scope. Additional text cleaning for other file formats requires significantly more implementation time.

- Data for knowledge base is accessible by all company employees in order to remove the risk of sensitive information leakage via the Q&A bot

- Availability of domain SMEs to collaborate with during the project span, especially for model evaluation

Challenges

- Building the right data set to fine-tune model

- Vector database integration (if not using Databricks Vector Search)

- Evaluating models, plus alleviating safety and bias concerns, is an active research community effort. There are currently no fully developed best practice standards.

Resources and schedule

HANDS-ON MVP

- Up to 16 data scientist days spread over 4 weeks, with 1 day of project management time

Evaluate Databricks for yourself. Visit us at databricks.com and try Databricks for free!