Summary

Large language models (LLMs) are the backbone of many natural language processing (NLP) applications, such as ChatGPT, Google Translate, etc. Most out-of-the-box LLMs are general-purpose models trained on publicly available text. However, many business problems need a specialized language model, 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 language model, such as Dolly
- 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 for chat, if you prefer not to purchase a license from OpenAI
  - Fine-tuning a model with your customized data for better embeddings if necessary, while leveraging compute best practices
  - Basic model evaluation
    - Any in-depth model evaluation requires manual inspection of results and collaboration with domain experts
  - Enabling internal teams
- 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 in-house model creation relative to paid services like OpenAI
- Reduce total cost of ownership (TCO)

Out of scope

- Configuration and integration of non-Databricks products
- Data cleansing associated with building broader data lake
- More than 100K documents with 500 words each
- Any efforts to improve an existing benchmark model has 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 a tabular format, ideally with these columns: title, content and web links. Web-scraping is out of scope. Additional text cleaning (e.g., for PDFs) 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

- 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