databricks

Service Description

Large Language Model (LLM) POC - Delivery Provider

This Service Description describes the scope of services (including associated Customer requirements) for the Databricks Advisory Service identified above (the "**Service**"), and applies to the Service under Customer's accepted Service Order.

Service Overview

Large language models (**LLMs**) are the backbone of many natural language processing (**NLP**) applications, such as ChatGPT. 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.

Leverage Retrieval-Augmented Generation (**RAG**) to enable the ability to answer questions based on your organization's knowledge base, integrated with your Databricks Lakehouse.

Objective

Develop a data strategy and custom LLM application to answer questions based on your organization's knowledge base, in your own Databricks environment.

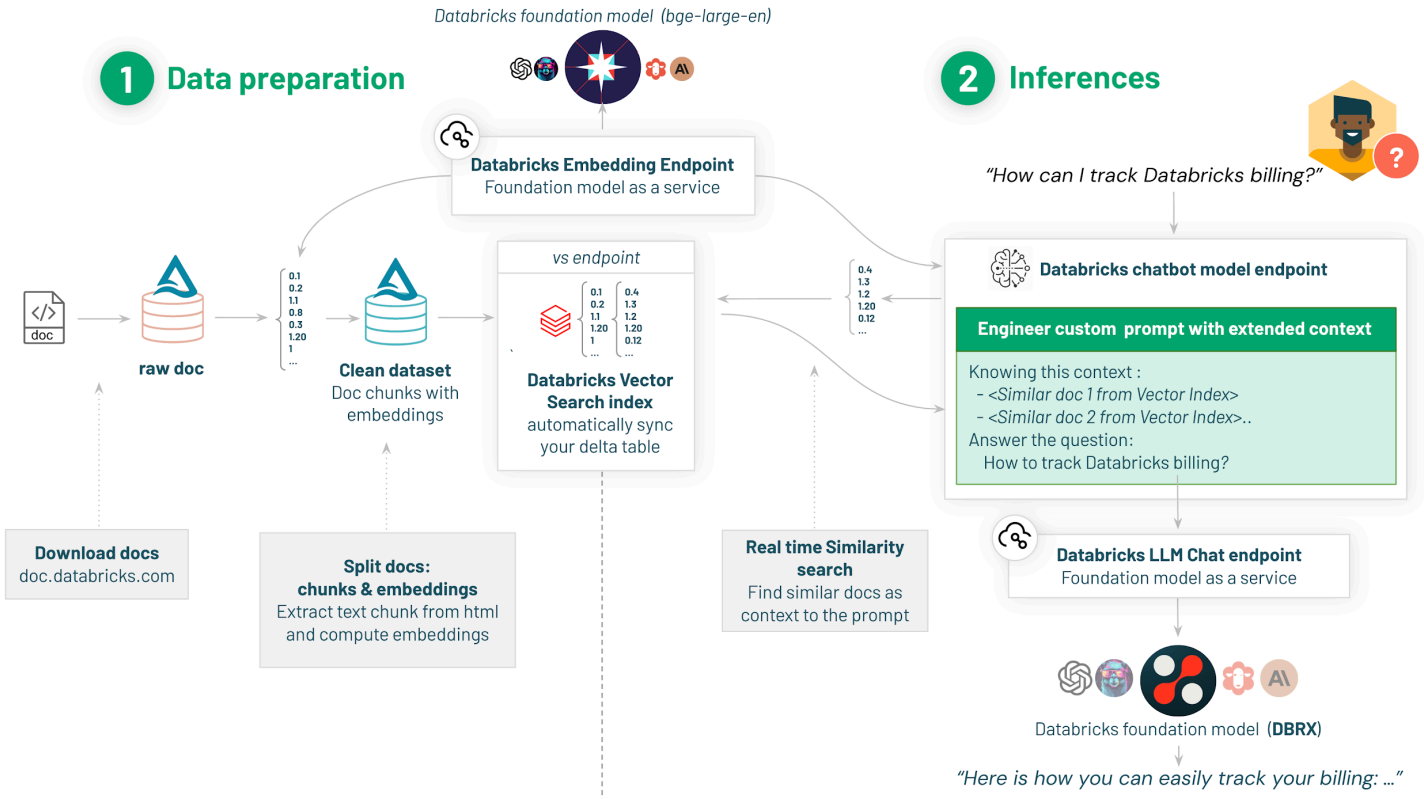
Description of Services

Databricks will provide Services from the Technical Focus Areas and Representative Activities described below in advising Customer on its LLM strategy and implementation approach. Specific activities performed will vary, depending on Customer-specific objectives. Databricks will deliver the Service through its subcontracted consulting provider, a member of the Databricks Delivery Provider Program ("**DPP**" or "**Delivery Provider**").

Technical Focus Area	Representative Activities
Advising *Advise on how to build your own dataset to get the most out of an LLM*	• Review techniques for implementing LLMs in your business including open source approaches and proprietary solutions • Advise on retrieval mechanism (e.g. vector stores) and data governance • Based on SLAs, licenses, etc. identify potential model families and sizes to consider for the generative component • Advise on end-to-end architecture options, from data ingestion to evaluation and deployment
Model Development & Evaluation	• Perform exploratory data analysis • Prepare data and create a vector store for retrieval

Technical Focus Area	Representative Activities
Build and evaluate a knowledge base Q/A model prototype	• Integrate an open source or proprietary LLM • Integrate components into an end-to-end prototype • Fine-tune a custom embedding model with your data (time permitting and if necessary) • Evaluate solution performance • Leverage Databricks Model Serving for real-time inference
Knowledge Transfer *Advise on next steps to deploy your Q/A model*	• Document the solution • Suggest next steps and best practices for Customer to plan solution deployment

The below visual illustrates a typical methodology and approach to these Services and the desired objective; note that this is an example, and not all steps / activities may apply to each Customer's project plan.



Prerequisites

Throughout the engagement, Customer will assure that the following requirements are met, to enable the Services:

- 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.
- Reasonable availability of Customer's appropriate technical, business, and domain experts to answer questions and provide necessary context, requirements, and information. These resources will be especially required during the model evaluation.
- Reasonable availability of Customer's technical resources familiar with Customer's relevant platform and data products, as well as those who will own such products after the Services engagement is complete for appropriate knowledge transfer to internal teams.
- Reasonable access to Customer environment, data, and artifacts as reasonably necessary for Databricks to successfully provide the Services on a timely basis.
- Customer will approve any models or services to be used in the engagement (including third party or open source models/services) and provide any required access keys if it wishes to use third party services (such as OpenAI).
- Additionally, if Customer does not leverage Databricks Vector Search for the retrieval component, Databricks recommends that data used in the solution be limited to only data that is accessible by all Customer personnel that have access to the solution (to reduce risk of unintended information being revealed via Q/A model output).

Out of Scope

- Web-scraping to obtain data, or preparing data that is not in the required format or type (see above).
- Removal of any personal data, confidential data, or other out-of-scope data from Customer-furnished data.
- Configuration of / integration with non-Databricks products, except as explicitly included in the Services activity description above.
- Fine-tuning a generative model is out of scope (if interested in this service, see GenAI Jumpstart).
- Multi-turn conversations are out of scope.
- Databricks cannot guarantee that any new model developed will outperform or improve upon an existing benchmark model.
- Integration of the Q/A bot with any front-end application.
- Developing a production-ready workflow (e.g., CI/CD).
- Any user acceptance testing, or any in-depth model evaluation that requires manual inspection of results or collaboration with domain experts.
- Training of end users.
- Work exceeding the allocation of Days and Services included in this engagement (see **Resources and Schedule**).

Resources and Schedule

Services consist of two **Days (16 hours)** of Project Management time, and up to **15.5 (fifteen and a half) Days** of Data Scientist/Machine Learning Engineer time, typically across a continuous 4-week period, applied against the Representative Activities in the Description of Services above.

Databricks will work with you to mutually agree to a project schedule as part of the Project Management phase. Resourcing assignments require a minimum 4-weeks advance request (while Databricks makes reasonable efforts to accommodate scheduling requests, personnel availability is subject to Databricks resourcing and discretion). Accordingly, Databricks recommends Customer coordinate with Databricks Services at least a month before placing its Service Order.

Additional Days of Data Scientist and/or Project Management time is available (by separate purchase) to apply to the activities outlined in this Service Description.

Additional Definitions and Terms

- "**Agreement**" means your agreement with Databricks providing general terms for our Services.
- **"Day"** means 8 working hours during local business days, excluding holidays.
- "**Services Order**" may be any of these mutually-accepted formats placed under your Agreement: an Order, Success Credit redemption request, written statement of work, or similar document
- "**we**", "**us**" or "**our**" means Databricks, Inc. or its Affiliates.
- "**you**" or "**your**" means the Customer organization that placed the Services Order