databricks

Service Description

GenAI Jumpstart

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

Accelerate your organization's ability to leverage proprietary data with Generative AI, through fine-tuning a large language model (LLM) on a specific business domain or task.

Objective

This engagement demonstrates the value of fine-tuning custom Generative AI models over leveraging out-of-the-box open-source models. It focuses on guidance in meeting your business goals by choosing optimal model architectures, data preparation methodologies, and other strategic decisions. The engagement also translates that strategy to implementation by training a reference fine-tuned LLM on your data with a reusable framework that can be applied to future fine-tuning tasks. This engagement will focus on ***one*** of the following fine-tuning approaches:

- **Supervised fine-tuning:** Train a model on structured prompt-response data. Used to adapt a model to a new task, change its response style, or add instruction-following capabilities.
- **Continued pre-training:** Train a model with additional text data. Used to add new knowledge to a model or focus a model on a specific domain.
- **Chat completion:** Train a model on chat logs between a user and an AI assistant. This format can be used both for actual chat logs, and as a standard format for question answering and conversational text. The text is automatically formatted into the appropriate chat format for the specific model.

Description of Services

Databricks will provide Services from the Technical Focus Areas and Representative Activities described below in assisting on building a reference Gen AI model. **Specific activities performed will vary, depending on Customer-specific objectives.**

Technical Focus Area	Representative Activities
Project Planning *Map out project plan to ensure success*	• Document and align on requirements of the business use case, success criteria, model architecture, preparation of proprietary data sets, evaluation strategies, etc. to ensure that the fine-tuning run will produce a quality model

Technical Focus Area	Representative Activities
Workload Setup *Implement a fine-tuned Gen AI model*	● Prepare data and transform it into the required structure and format for fine-tuning ● Benchmark out-of-the-box open source models using evaluation criteria defined by Customer ● Execute a fine-tuning run to train the reference model, capturing parameters, metrics and model artifacts to MLflow ● Liaise with the Generative AI Applied Research team if additional support is needed
Model Serving and Future Work *Serve fine-tuned model as REST API & next steps*	● Conduct model evaluation and interpret model results ● Stand up Databricks Model Serving endpoint with fine-tuned model, making the model available as a REST API ● Provide guidance on next steps (e.g., pretraining, scaling fine-tuning, RAG, etc)

Prerequisites

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

- Customer has knowledge or experience with open-source models with a well-defined use case.
- Customer to provide access to training data to build the reference model. If Customer does not provide access, Customer will identify publicly accessible, representative datasets to cover Customer's stated use case.
- For instruction fine-tuning (i.e., adapting an LLM to a specific task), Customer must provide an instruction dataset. This dataset should consist of prompt-response or chat-formatted examples which will be used to fine-tune the LLM. Thousands of examples are recommended.
- Customer must have pre-defined criteria to evaluate the resulting fine-tuned model.
- Fine-tuning the reference model will be limited to supported models.
- Customer must provide a Databricks workspace, even if Customer is using Multi-Cloud Training (also called, MCT or MosaicML) as it is required for data preparation.
- 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 Provider to successfully provide the Services on a timely basis.

Out of Scope

- Development outside of Databricks or MosaicML (also called, Multi-Cloud Training or MCT) (*see* Prerequisites above).
- Assistance during a Hero Run (if interested in this service, *see* Hero Run Enablement Services).
- Developing a production-ready workflow involving CI/CD, infrastructure-as-code, etc. (though advising is in scope).
- Prompt engineering strategy (though advising is in scope).
- Work exceeding the allocation of Days and Services included in this engagement (see **Resources and Schedule**).
- The compute cost for fine-tuning the model is NOT included.

Resources and Schedule

Services consist of **up to 15 Days** of Data Scientist time, and up to **2 Days** of Project Management time, typically across a continuous **3-4** 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