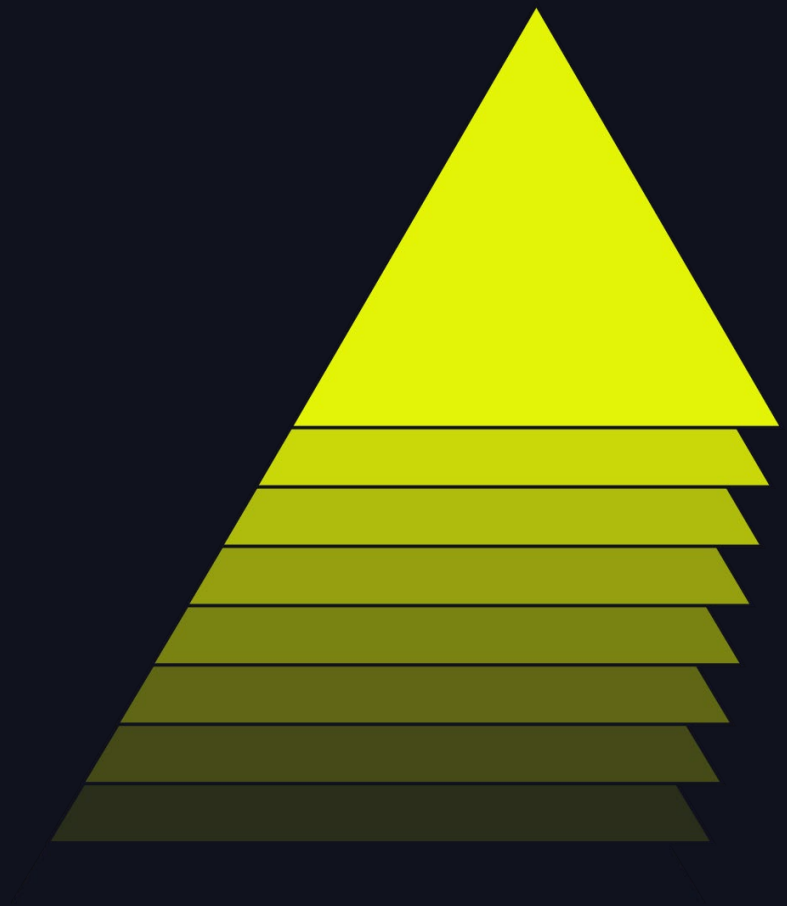
DATA+AI SUMMIT
BY databricks

Advanced Experiment Tracking for LLM-Powered Applications with Customized Open-Source MLflow



Manas Mukherjee, Sr. Staff Software Engineer, Intuit
Date: 06/11/2024

Agenda

LLM powered app evaluations tracking using MLFlow

- LLM Applications Overview
- Why Open Source MLflow
- Evaluation Process
- Tracking & Selection of Run & Review
- Next Steps

Introduction to LLM-Powered Applications

- **LLM Applications at Intuit**
  - Intuit Assist: Guides small business owners in QuickBooks
  - Creates launch announcements for marketers in Mailchimp
  - Helps Credit Karma members make smart money decisions
  - Transforms tax preparation in TurboTax
- **Importance of Experimentation**
  - Tailoring prompts and configurations for diverse use cases
  - Ensuring high-quality, reliable AI-driven solutions
  - Continuous improvement through iterative testing

Why open source MLflow ?

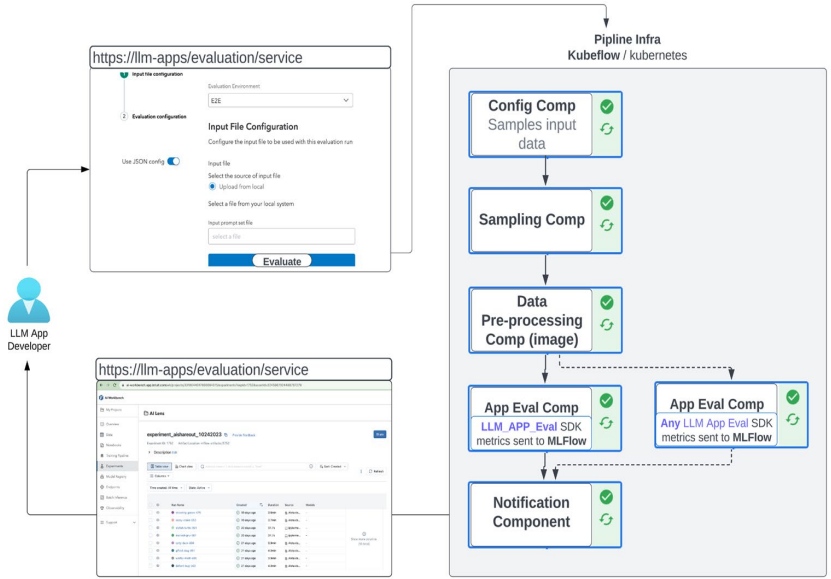
- **Flexibility Across Cloud Platforms & Tools**
  - Supports various cloud tools: notebooks, spark, pipelines etc.
  - Enables users to stay within their preferred cloud solutions and workflows
  - Offers a multi-tenant solution with an intuitive internal UX
- **Self-managed backend data**
  - Seamless integration with internal databases and integrations with others MDLC components.

```
mlflow server \
  --backend-store-uri postgresql://$DB_USER:"$DB_PWD"@$DB_HOST:$PORT/$DB_NAME \
  --artifacts-destination $MLFLOW_ARTIFACTS_BUCKET_NAME
```

- **Leverage robust open-source capabilities**
  - Accelerate development process
  - Contribute back to the open-source community

LLM App Evaluation Workflow

Scalable Evaluation Pipeline & Experiment Tracking using MLFlow



1. Users vary prompts, parameters, and guardrails to start experiments.
2. Benchmark Datasets test model performance with and without ground truth responses.
3. Advanced LLMs perform detailed evaluations using Custom Rubrics.
4. Standard Metrics like BLEU scores and BERT scores measure hallucinations and answer relevance.
5. Batch Jobs execute evaluations using Kubeflow/Kubernetes.
6. Results are captured and logged in MLflow, including metrics, tags, and artifacts.

GenAI App Evaluation using MLFlow

App evaluation component uses eval data and config

Evaluation Setup (data, config)

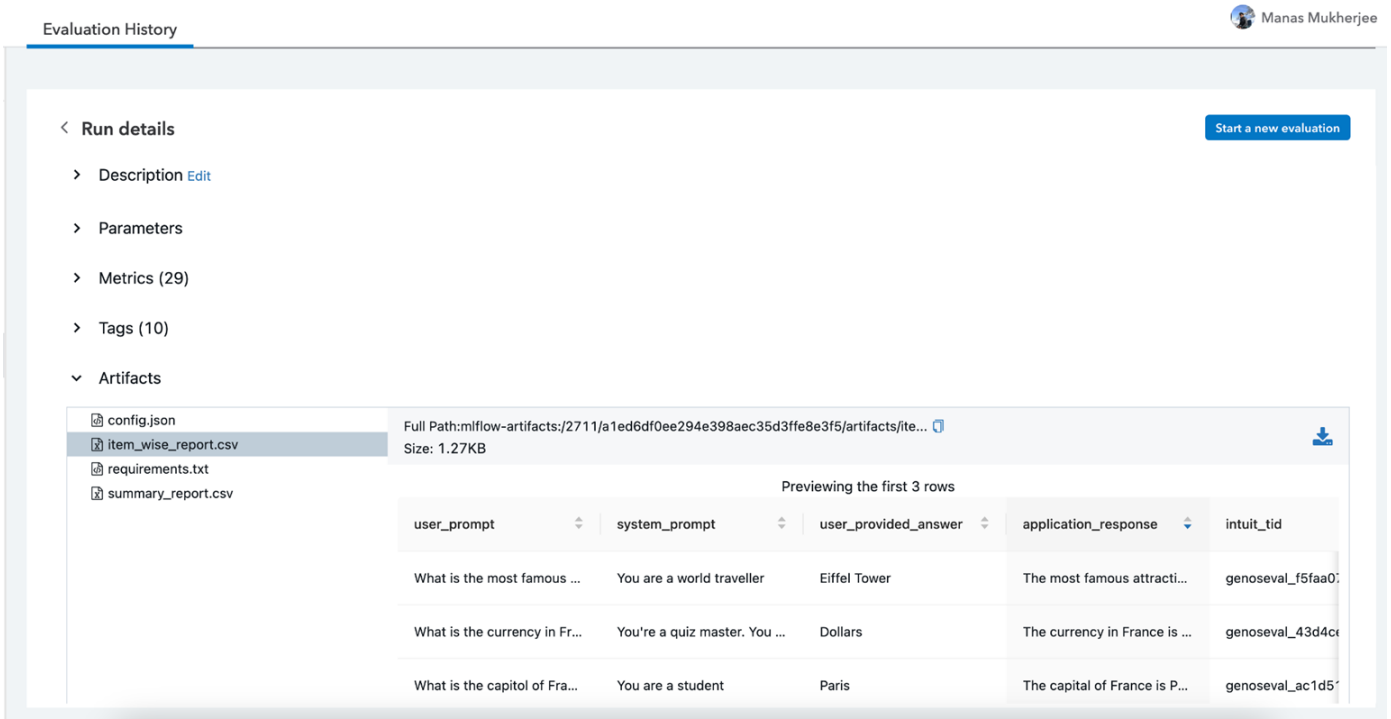
```
# 1. Evaluation Data

evaluation_data = [user_prompt, system_prompt_ref, expected_answer]

# 2. Evaluation Config

eval_config = {
   "metrics" : [
                "AutoEval WithOut Ground Truth",
                "RubricBasedLLMEval With Ground Truth",
                ...other metrics
              ],
   "mlflow_exp_config" : {
       "use_case_identifier": "genai_qbo_prd",
       "exp_name": "app1_eval_prompts_june11"
   }
}
```

Llm evaluation component

```
# LLM powered App Evaluation
from llm_app_eval import evaluate

mlflow.start_run(experiment_id=eval_config.mlflow_exp_config.exp_name):
    ## Evaluation using llm_app_eval (or any other) library/methods

    evaluation_instance = evaluate.Evaluate(config)
    reports = evaluation_instance.generate_reports(df)

    ## Log evaluation result using MLFlow

    mlflow.log_metrics(reports.metrics)
    mlflow.log_tags(reports.tags)
    mlflow.log_artifacts(reports.artifacts)
```

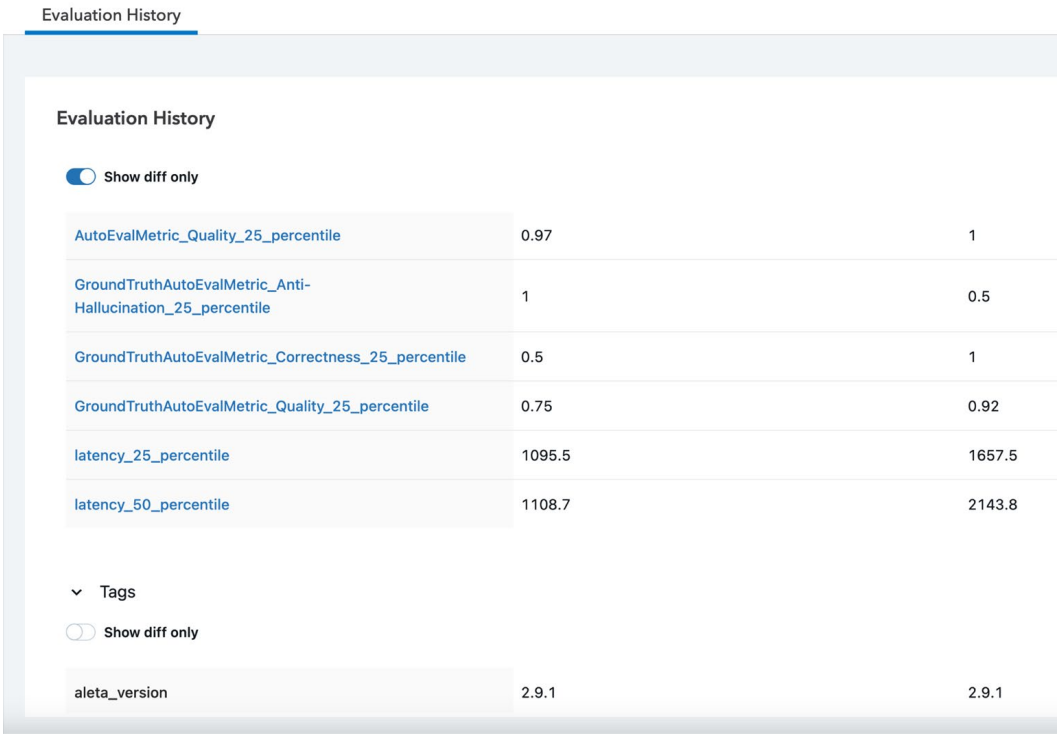
MLFlow: Tracking Eval Runs

Track: Aggregate metrics, tags, artifacts



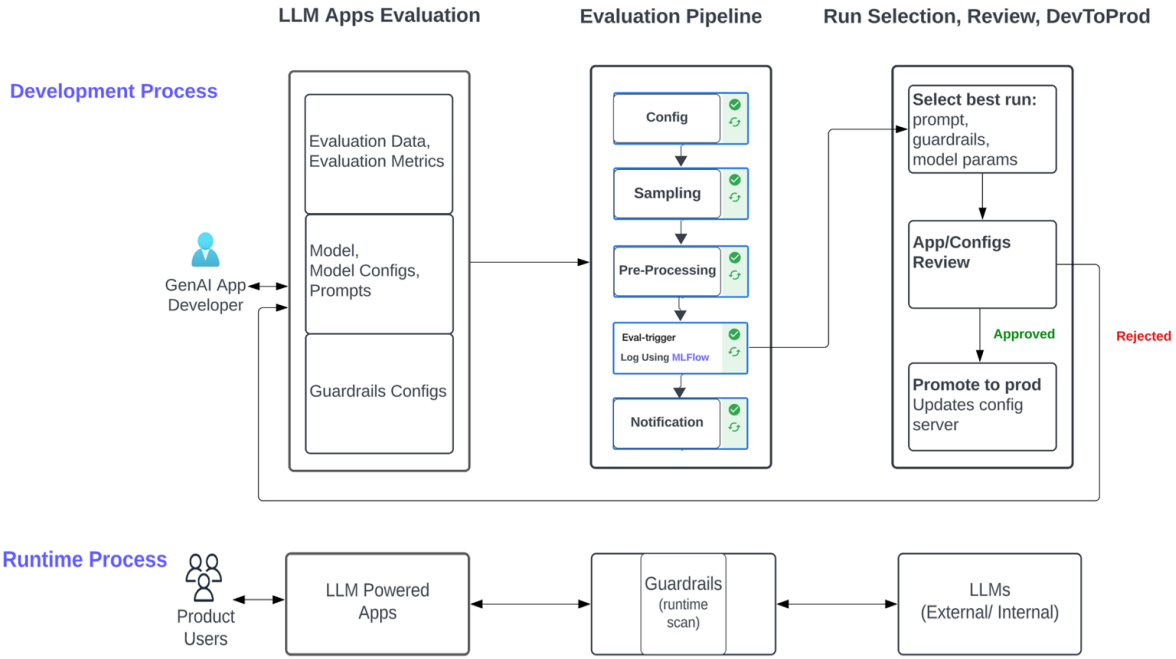
Comparing and Selecting Best Run

Compare: Aggregate metrics, prompts, model params, guardrail -settings



Experimentation to Production Journey

LLM powered App Lifecycle



Next Steps: LLM App Evaluation

At Intuit

- ❏ New MLFlow Capabilities
  - ❏ Mlflow llm evaluation: `mlflow.evaluate()`
  - ❏ App/Domain specific evaluation dataset: `mlflow.data.dataset.Dataset()`
- ❏ Enhancement Proposals
  - ❏ Need for UI Plugins
    To avoid reverse engineering, and easy integration with custom solution
  - ❏ Experiment Tagging
    To use MLflow as a multi-tenant solution
  - ❏ LLM as a judge - `mlflow.metrics.genai.make_genai_metric()`
    Use custom proxy server (without using MLflow Deployments for LLMs)

Thank You

- Databricks
- MLFlow