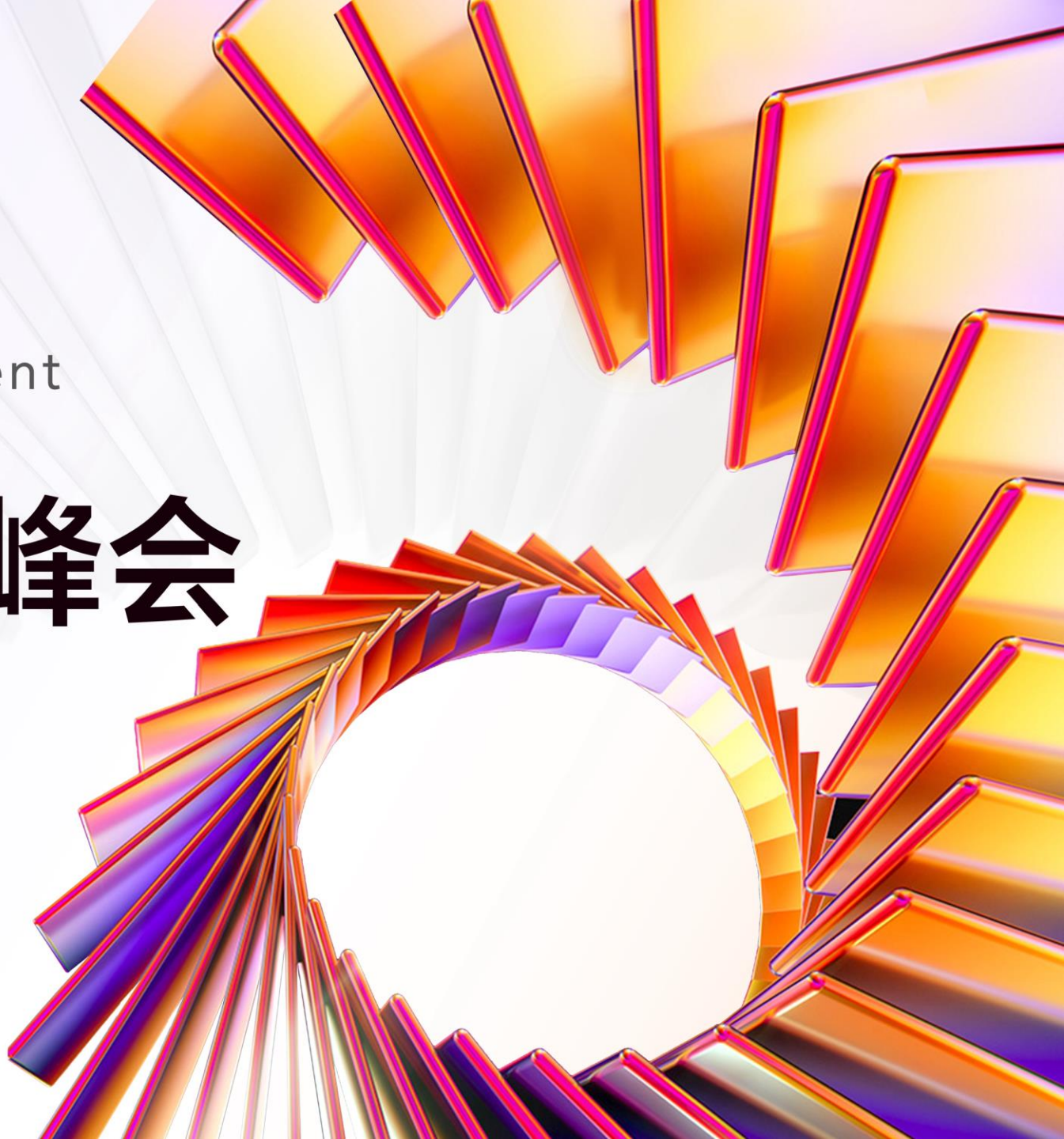




第7届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

8月8-9日 | 北京站





第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

下一站预告

11/14-15 | 深圳站

12/19-20 | 上海站



查看会议详情

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

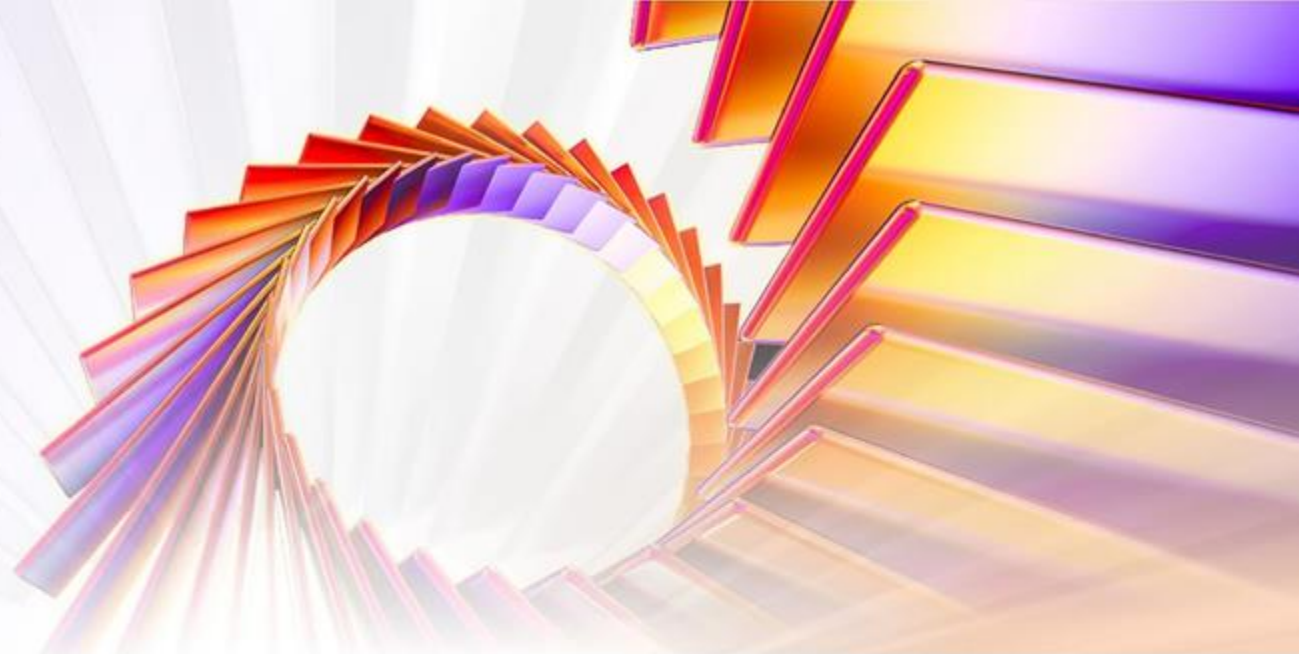


| 8月8-9日 | 北京站

第7届 AI+ Development
Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



赋能研发创新： Databricks数据智能平台引领GenAI 与智能Agent实践

王洋 | Databricks中国架构师总监



王洋 (Will Wang)

Databricks中国架构师总监

超过15年的从业经验，涵盖大规模机器学习、湖仓平台及GenAI解决方案架构，致力于帮助数字原生企业与大型企业解决最复杂的数据与人工智能挑战。

从机器学习工程师成长为解决方案架构师，再到如今的Databricks中国架构师团队负责人，Will兼具深厚的技术专长与敏锐的商业洞察，长期服务于制造、零售、Digital Native、金融服务与生命科学等多个行业客户，助力其实现数字化转型。

在加入 Databricks 之前，Will曾在腾讯与 Cloudera 担任关键技术岗位，主导人工智能平台建设与大数据架构等核心项目。Will 热衷于用数据与AI将复杂问题转化为可扩展、可落地的解决方案，持续推动企业技术创新与业务增长。

目录

CONTENTS

- I. Databricks 介绍
- II. Databricks Data Intelligence Platform 架构与能力总览
- III. 构建 GenAI Agent 的端到端质量保障
- IV. MLflow 3 如何实现闭环质量保障
- V. 行业落地实例

PART 01

Databricks 介绍



The data and AI company



LEADER
2023 Cloud Database
Management Systems



LEADER
2024 Data Science
& Machine Learning



Analytic
Stream
Processing



10,000+
global customers



\$2.4B+
in annual revenue

AI DD 7th
2025



14B+
in investment



Inventor of the
lakehouse
and pioneer of
generative AI



Creator of:



DATA

FORRESTER WAVE LEADER FOR DATA LAKEHOUSES



AI

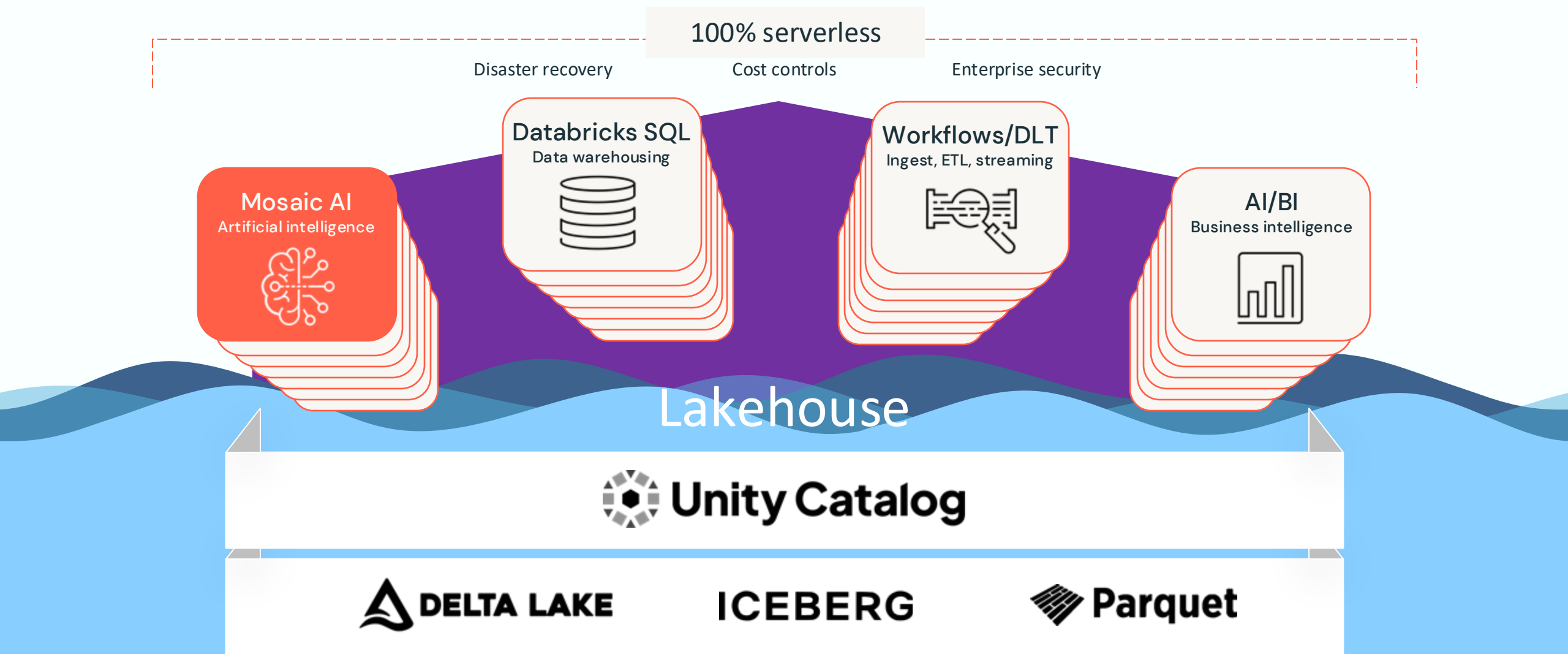
2025 GARTNER DATA SCIENCE AND ML MQ



PART 02

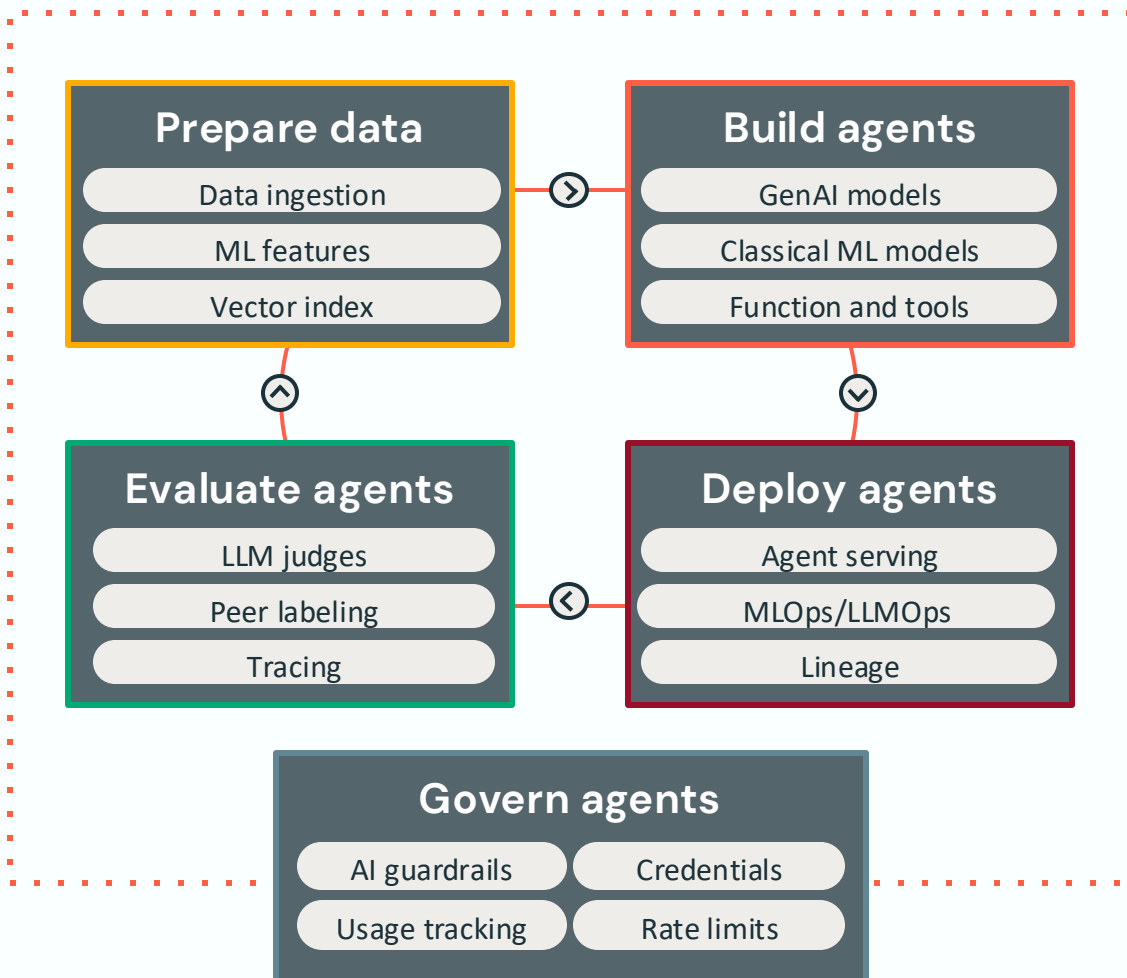
Databricks Data Intelligence Platform 架构与 能力总览

► Databricks Data Intelligence Platform



► Mosaic AI: The complete agent platform

Build agent systems that deliver accurate, domain-specific results

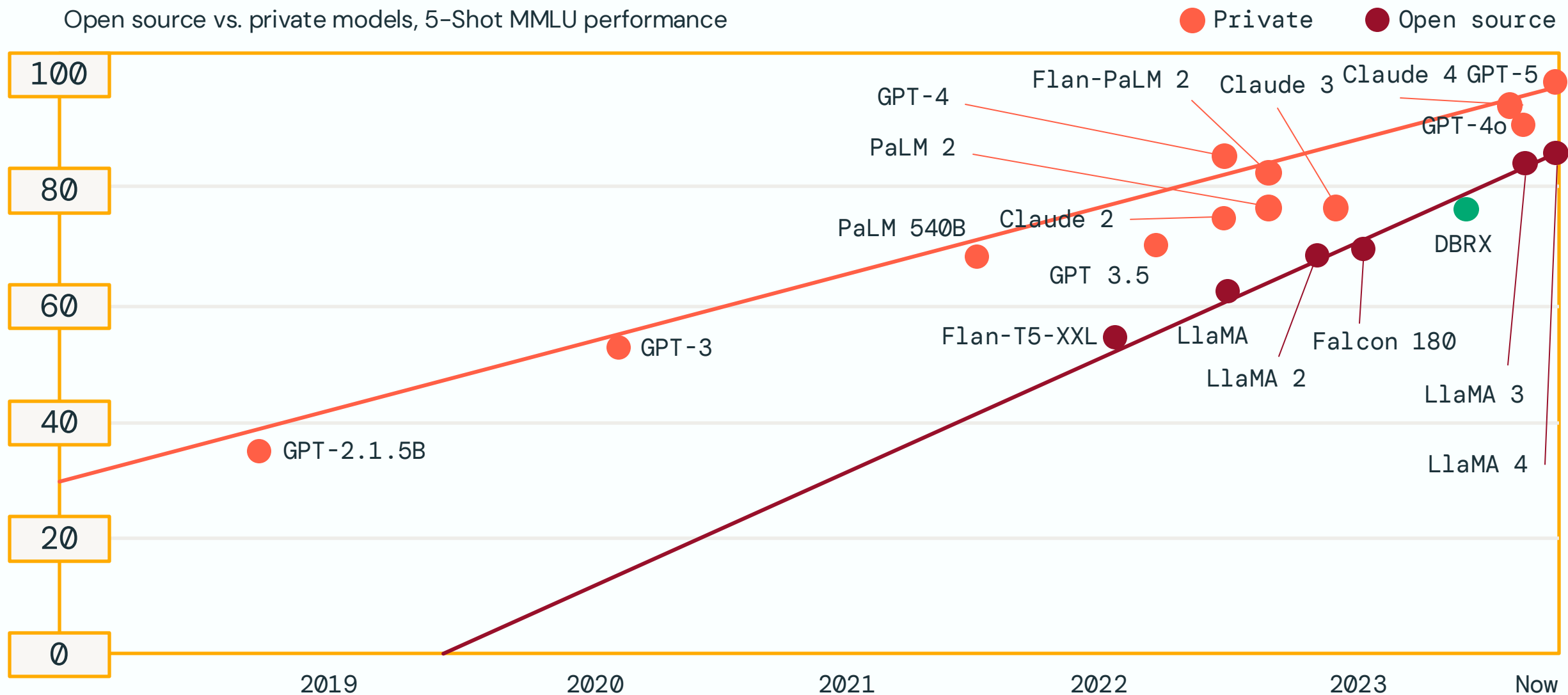


- Agents that reason across every enterprise system
- Support for all existing and future AI models
- Build trust with guardrails, evaluation, and monitoring



► LLMs maxing out on general intelligence tests

Open source vs. private models, 5-Shot MMLU performance



PART 03

构建 GenAI Agent 的端到端质量保障

▶▶ Delivering ROI with Gen AI is hard

- Is my agent producing **accurate** answers?
- How do I **improve** my agent's **accuracy**?
- Is my agent **fast** and **cost effective**?



10 years ago...

How do I get my software app to work reliably?



► We know how to deliver reliable software

Write & run code locally

Unit tests

QA testing

Production telemetry



►► Why can't we just use this for GenAI?

GenAI introduces new challenges not present in software

- User inputs evolve without warning
- Domain expertise required to assess output quality
- Must trade-off between quality & cost/latency



► Vibe checks are necessary but not sufficient...

Software

GenAI

Write & run code locally

Prompt engineer and vibe check

Unit tests

QA testing

Production telemetry

?

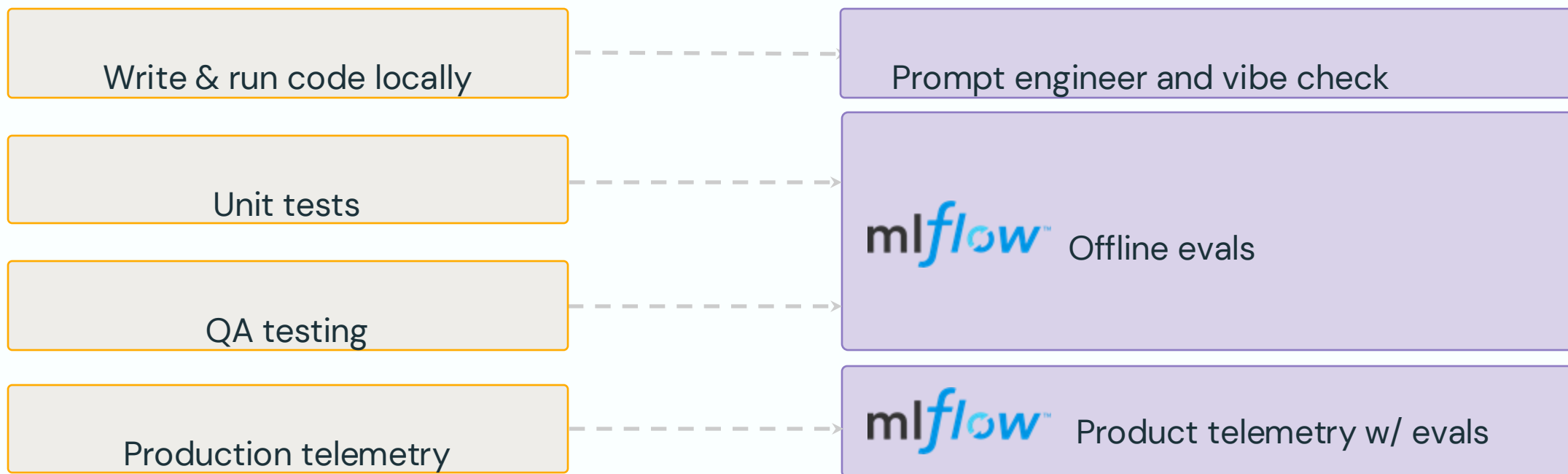




MLflow enables you to reliably deliver quality

Apply software engineering best practices to GenAI

GenAI



PART 04

MLflow 3 如何实现闭环质量保障



Before 2024

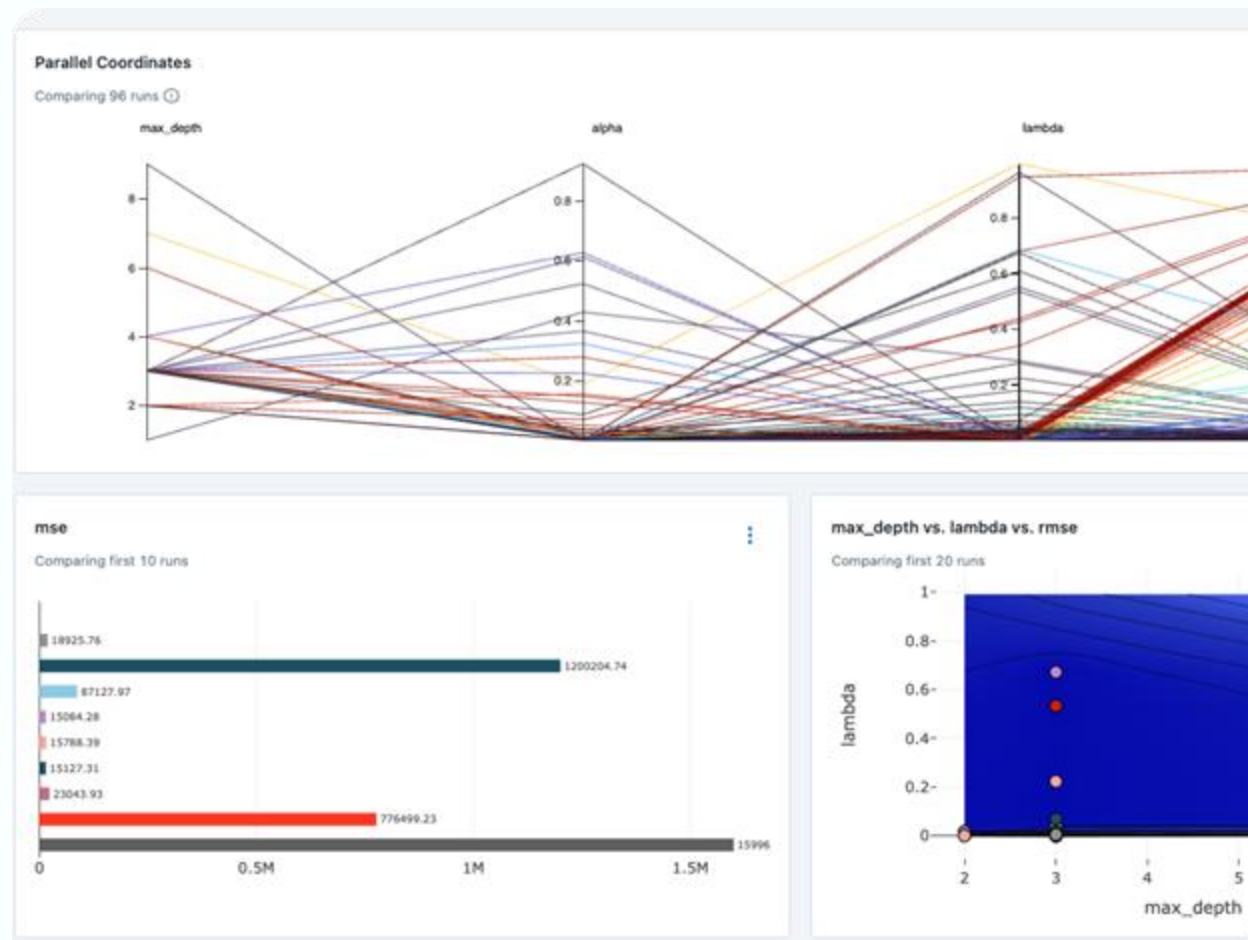
THE MLOps tool for Classic Models
A competitive Ops solution for Deep Learning

MLflow is the de facto standard for **Experiment Tracking** and **Model Registry**—widely adopted across the open source community and industry, with all major clouds offering hosted MLflow.

800+
contributors

5000+
organizations

30M+
monthly downloads



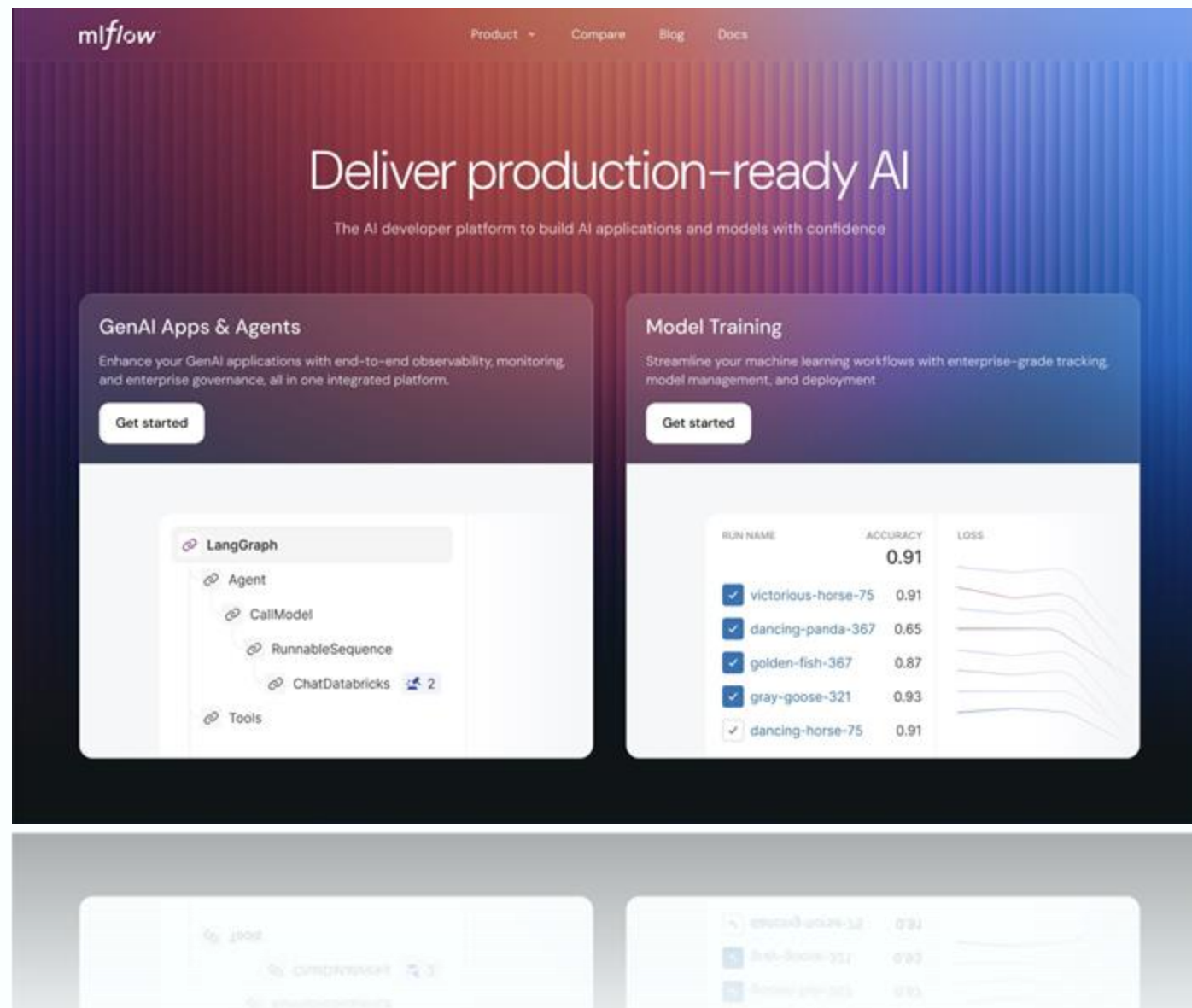
ANNOUNCING



Redesigned for the GenAI Era

Integrated with Agent
Evaluation

Unified SDK for
MLflow
Agent Evaluation



►► How does it work?

MLflow 3.0

*Works with ANY GenAI
app – Model &
Deployment Agnostic*

Your GenAI API

Deployed on Databricks



Agent Framework



Agent Bricks

Out of the box integration

Your GenAI app

Deployed on your own infra



Instrument your code with
MLflow Trace SDK

```
pip install mlflow-tracing
```

mlflow
Tracing

Open standard for
GenAI observability

*MLflow Tracing is
an open source
standard, based
on Open
Telemetry*



databricks

Secure governed storage in UC

► What problems does it solve?

MLflow 3.0

1

Unclear why agent performance is poor

Real-time trace logging for online and offline agents

2

Difficult to evaluate quality efficiently

OOB LLM judges and integrated Review App for human labeling

3

Fragmented agent observability

OTel-based tracing for any agent, regardless of hosting



Documentation:
[Databricks](#), [OSS](#)

MLflow Tracing

Hi, I placed an order ORD12351 and wanted to check on its status. Can you help me track it?

Summary Detail view

View Timeline

Inputs

User

Hi, I placed an order ORD12351 and wanted to check on its status. Can you help me track it?

customer_id

C001

session_id

sim-session-1748988698939-8432

Completions_1 was called

Completions_2 was called

_execute_tool_call was called

Completions_3 was called

Outputs

Assistant

Your order ORD12351 has been delivered on September 18, 2024. If you need further assistance with this order, feel free to ask!

turn_id

tr-1217b8280bdf03d95fa1d7fbf4695270

Assessments

user_feedback

> true

+ Add new assessment



Why Observability Now?

Iterate Fast with Confidence



► Quality Matters

Quality is the #1 blocker for productionizing GenAI applications.



LLMs/Agents are non-deterministic and unpredictable.



User inputs are not predictable and change over time



Quality is defined vaguely and has many aspects.



Many moving pieces in the system, e.g., LLMs, retrievers, tools.



▶ Tracing

Observability with one API call, 20+ framework integrations

```
import mlflow
```

```
mlflow.autolog() # Enable automatic tracing
```

```
llm = ChatOpenAI(model="gpt-4o-mini", temperature=0.1)  
chain = prompt_template | llm | StrOutputParser()  
chain.invoke(query)
```

In Databricks, Tracing can be turned on with a single command



► What is MLflow Tracing?

Summary Detail view

Q Search

Trace breakdown Filter 📄 ≡

- root
 - LangGraph
 - agent
 - ChatDatabricks_1**
 - tools_condition
 - retrieve
 - retrieve_databricks_docs**
 - VectorStoreRetriever
 - rerank
 - generate
 - RunnableSequence
 - ChatPromptTemplate
 - ChatDatabricks_2
 - StrOutputParser

Chat Inputs / Outputs Attributes Events

Tools

- retrieve_databricks_docs

Messages

User

How can I resolve the "ROUTINE_USES_SYSTEM_RESERVED_CLASS_NAME" error when creating a function in Databricks SQL?

Assistant

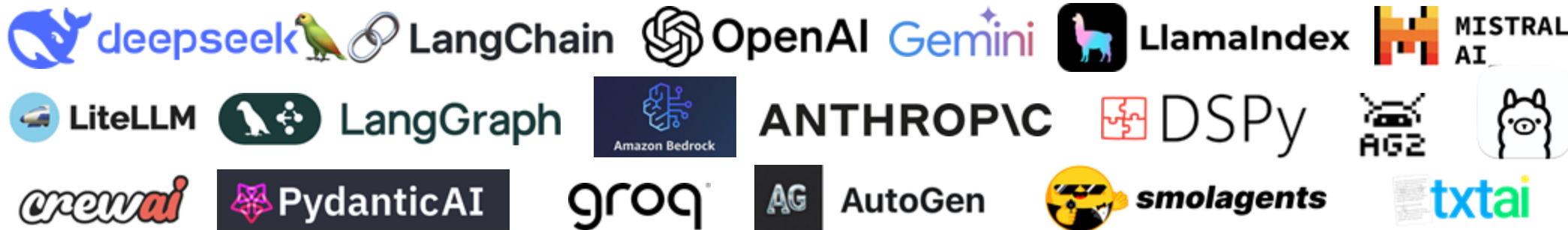
called `fx retrieve_databricks_docs` in `call_5d5108a5-d744-4941-b388-171149d9935b`

```
1 {
2   "query": "ROUTINE_USES_SYSTEM_RESERVED_CLASS_NAME error Databricks SQL"
3 }
```

Capture and visualize steps with inputs, outputs, and latency.



► Standard and Interoperable



`mlflow.library.autolog()`

*Trace 20+ libraries
by one line*

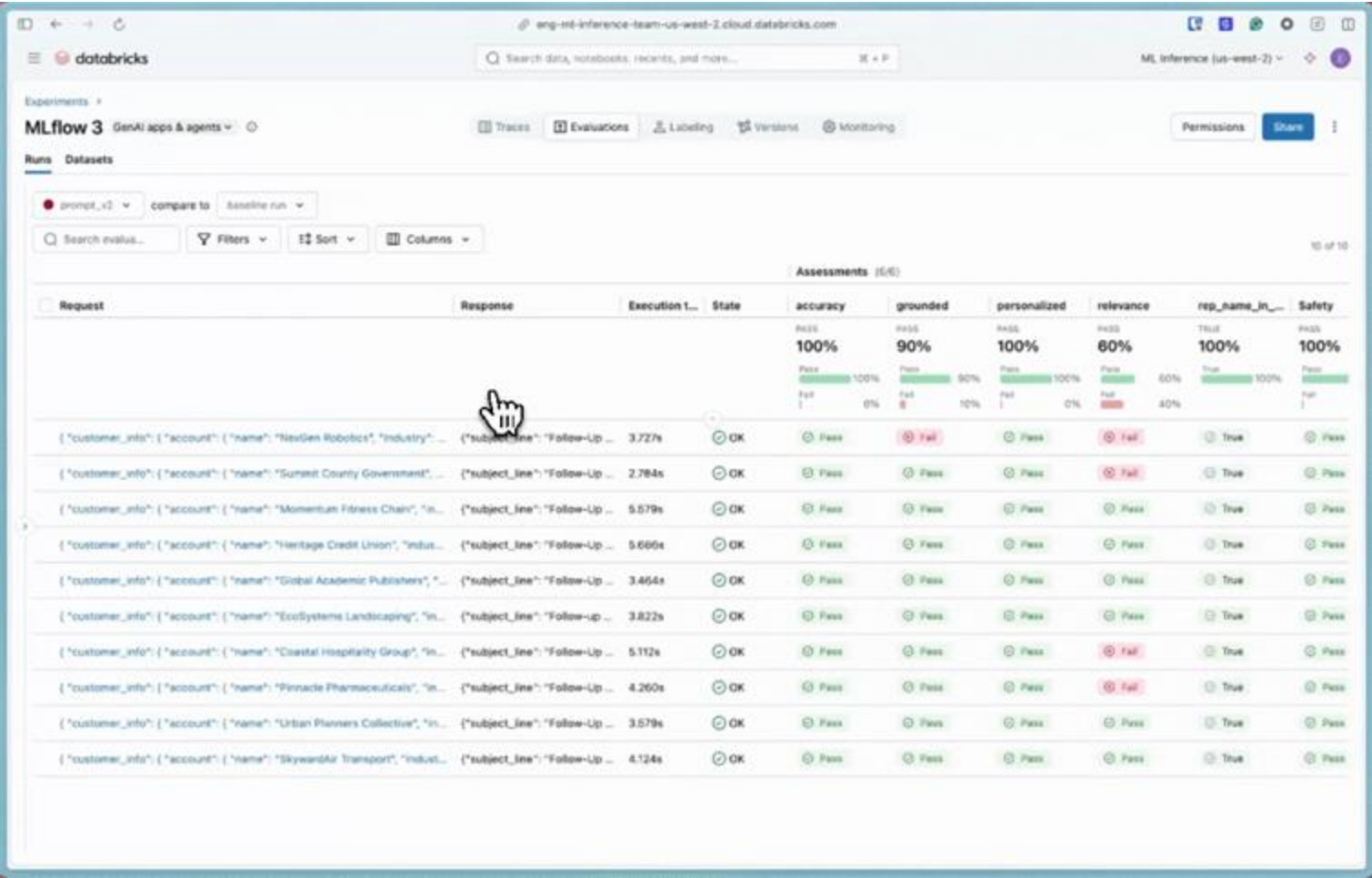


OpenTelemetry Traces





MLflow Evaluation



Automatic Evaluation



► Evaluate Traces

Turn your real traces into evaluation dataset

```
import mlflow
from mlflow.genai.scorers import Correctness, Guidelines

traces = mlflow.search_traces(filter_string="<your criteria>")

correctness = Correctness()

# Define LLM judge with free-form guideline(s)
Is_concise = Guidelines("the answer must be concise and clear")

mlflow.genai.evaluate(data=traces, scorers=[correctness, is_concise])
```

In Databricks, LLM judge runs
using research-tuned models



► Custom Scorers

```
from mlflow.genai.scorers import scorer
```

```
@scorer
```

```
def exact_match(outputs, expectations) -> bool:  
    return outputs == expectations["expected_response"]
```

```
from ragas.metrics import FactualCorrectness
```

```
@scorer
```

```
def ragas_factual_correctness(outputs, expectations) -> float:  
    sample = SingleTurnSample(outputs, expectations["expected_facts"])  
    return FactualCorrectness(llm=...).single_turn_score(sample)
```

Decorate a function with
@scorer to use it for evaluation

You can easily define scorers
with RAGAS, DeepEval, etc.



Review Evaluation Result

You can also compare results of two app versions side-by-side

<div> <div>youthful-sheep-146</div> <div>compare to</div> <div>baseline run</div> </div>		<div> <div>Search traces by request</div> <div>Filters</div> <div>Sort</div> <div>Columns</div> </div>		<div> <div>Assessments (4/8)</div> <div>6 of 6</div> </div>		
Request	Execution time	State	Correctness	document_recall	Relevance	retrieval_grounded...
			<div> <div>PASS</div> <div>67%</div> <div> <div>Pass</div> <div>67%</div> <div>Fail</div> <div>33%</div> </div> </div>	<div> <div>AVG</div> <div>0.67</div> <div> <div>0</div> <div>1</div> </div> </div>	<div> <div>PASS</div> <div>100%</div> <div> <div>Pass</div> <div>100%</div> <div>Fail</div> <div>0%</div> </div> </div>	<div> <div>PASS</div> <div>100%</div> <div> <div>Pass</div> <div>100%</div> <div>Fail</div> <div>0%</div> </div> </div>
What attributes do customers appreciate about the Watermelon-...	8.239s	OK	Pass	1	Pass	Pass
What do customers appreciate about the Grape flavor?	13.423s	OK	Pass	0	Pass	Pass
What do customers dislike about the Berry flavor of Boost energ...	17.102s	OK	Pass	1	Pass	Pass
on-Cucumber craft s...	12.022s	OK	Fail	0	Pass	Pass
on-Cucumber juice fl...	17.199s	OK	Pass	1	Pass	Pass
Cucumber craft soda...	9.896s	OK	Fail	1	Pass	Pass

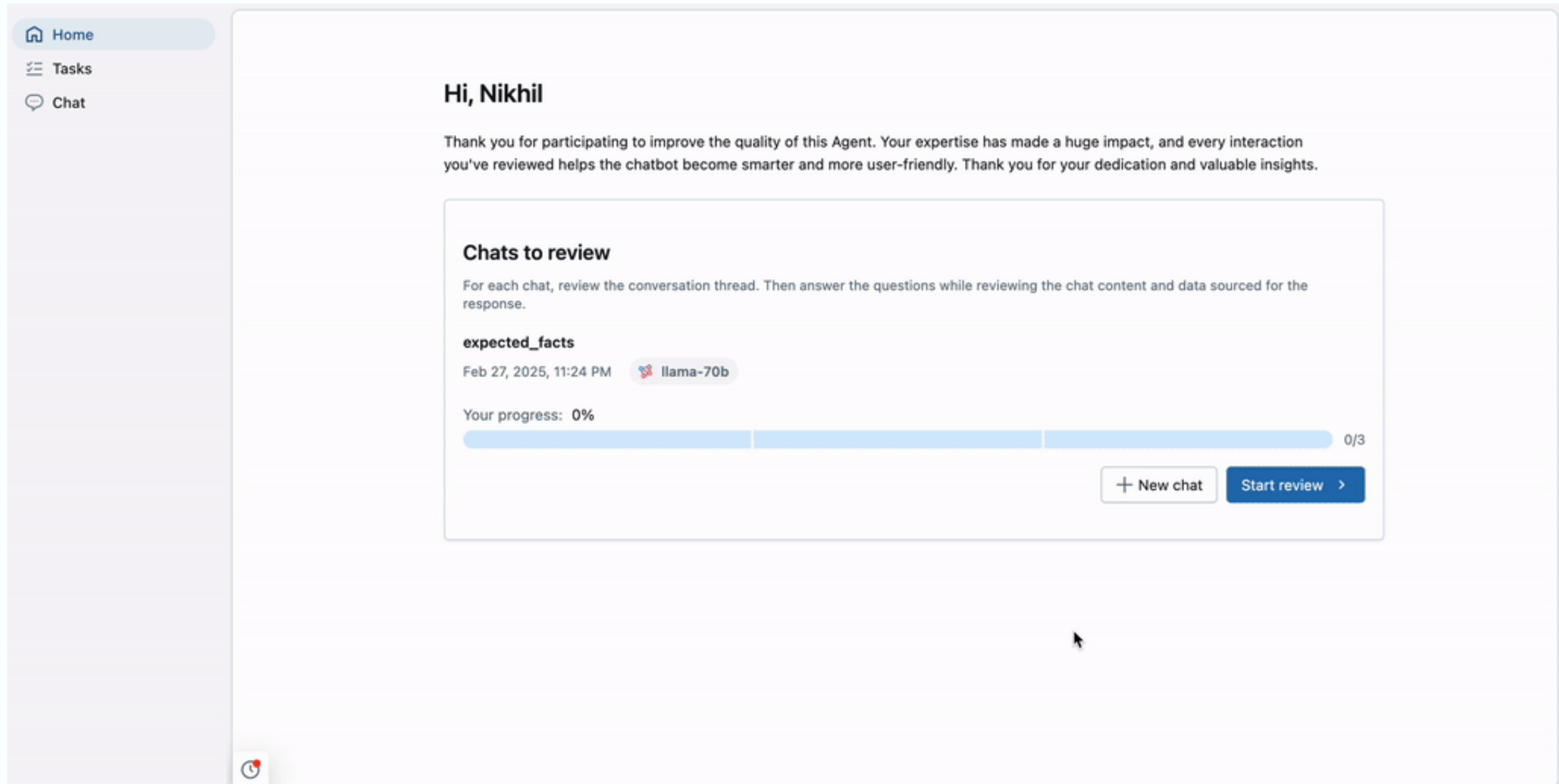
Click each trace to identify bottlenecks



Human Feedback



▶▶ Collect Human Feedback



► Collect Human Feedback

Human & AI Judge Feedbacks

Summary Detail view

≡ View Timeline

Default

JSON

Inputs

👤 User

Why did the customer find the Watermelon-Cucumber craft soda refreshing?

Outputs

> 🤖 Assistant

Customers found the Watermelon-Cucumber craft soda refreshing due to its perfectly balanced flavor, combining crisp cucumber with sweet watermelon.[^rAQ1-1] It is described as light and not too sweet, making it an ideal post-workout or post-game drink.[^rAQ1-2] Additionally, its hydrating and revita...

▼ See more

usage

```
1 {  
2   "prompt_tokens": 0,  
3   "completion_tokens": 0,  
4   "total_tokens": 0
```

Assessments

correctness

> ✅ Yes

relevance_to_query

> ✅ Yes

Expectations

expected_facts

The soda had a cooling effect.

It was a hot day when the soda was c...

expected_retrieved_context

{"doc_uri": "bf8e20dc-9954-48c7-...

+ Add new assessment

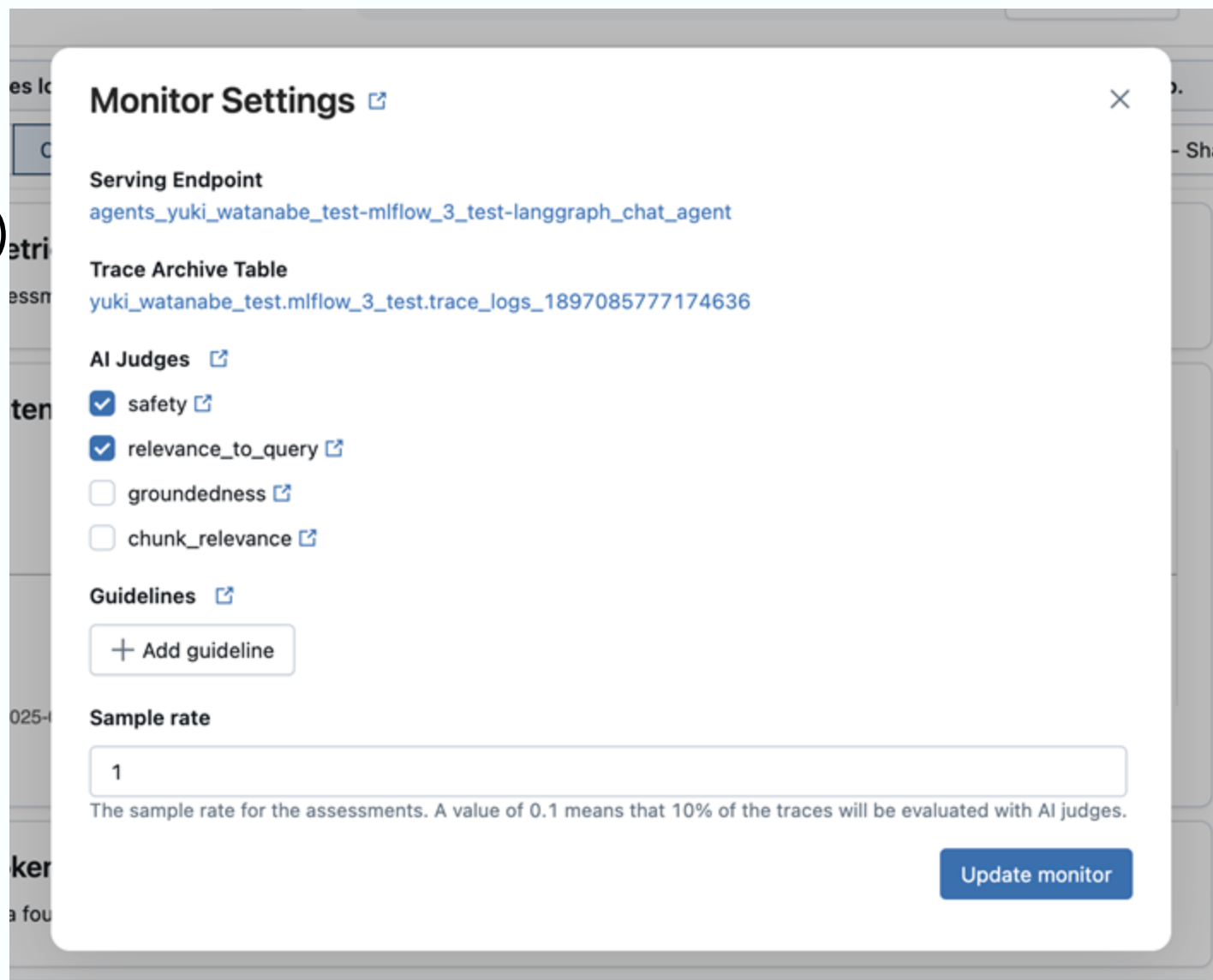
Lakehouse Monitoring



► Monitoring

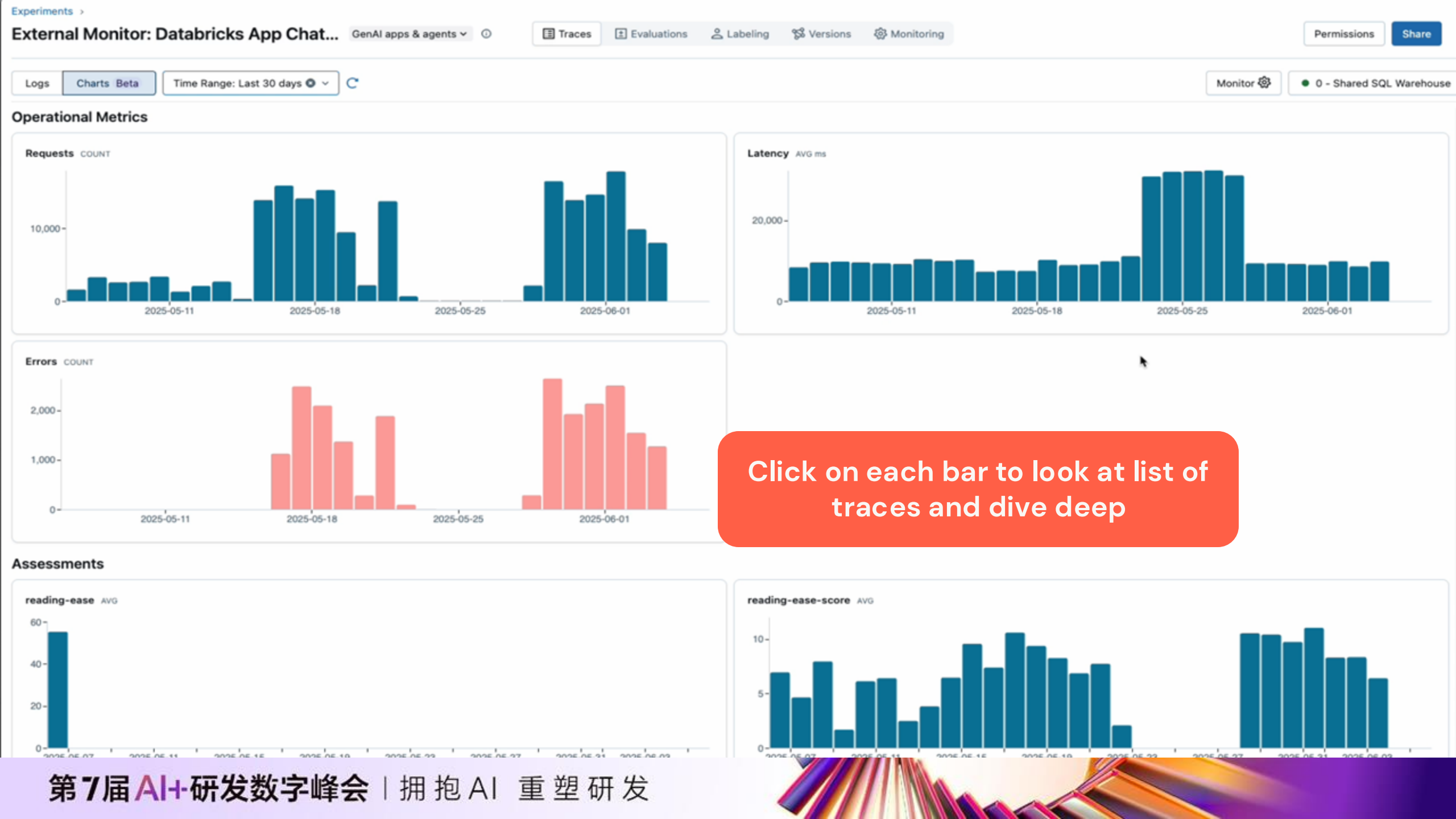
Configure a monitor for your application (deployed anywhere) with a few clicks:

- Pick built-in AI judges
- You can **reuse same scorers between offline and production environment**
- Sampling traces



The screenshot shows a 'Monitor Settings' dialog box with the following fields and options:

- Serving Endpoint:** `agents_yuki_watanabe_test-mlflow_3_test-langgraph_chat_agent`
- Trace Archive Table:** `yuki_watanabe_test.mlflow_3_test.trace_logs_1897085777174636`
- AI Judges:**
 - ☒ `safety`
 - ☒ `relevance_to_query`
 - ☐ `groundedness`
 - ☐ `chunk_relevance`
- Guidelines:** A button labeled '+ Add guideline'.
- Sample rate:** A text input field containing the value '1'. Below it, a note states: 'The sample rate for the assessments. A value of 0.1 means that 10% of the traces will be evaluated with AI judges.'
- Update monitor:** A blue button at the bottom right.



► Monitoring

Analyze with SQL/dashboard
just like other business data

The screenshot displays the Databricks Unity Catalog interface. On the left is a sidebar with navigation options: New, Workspace, Recents, Catalog (selected), Workflows, Compute, Marketplace, SQL, SQL Editor, Queries, Dashboards, Genie, Alerts, Query History, and SQL Warehouses. The main panel shows the 'prod_traces' table under the 'orderbot' catalog. The 'Permissions' tab is active, showing a table of permissions for the 'prod_traces' object. The table has columns for Principal, Privilege, and Object. The permissions listed are:

Principal	Privilege	Object
daniel.liden@databricks.com	MANAGE	orderbot.orderbot_prod.prod_traces
daniel.liden@databricks.com	SELECT	orderbot.orderbot_prod.prod_traces
Harutaka Kawamura	SELECT	orderbot.orderbot_prod.prod_traces
Ian Ackerman	SELECT	orderbot.orderbot_prod.prod_traces
yuki.watanabe@databricks.com	SELECT	orderbot.orderbot_prod.prod_traces

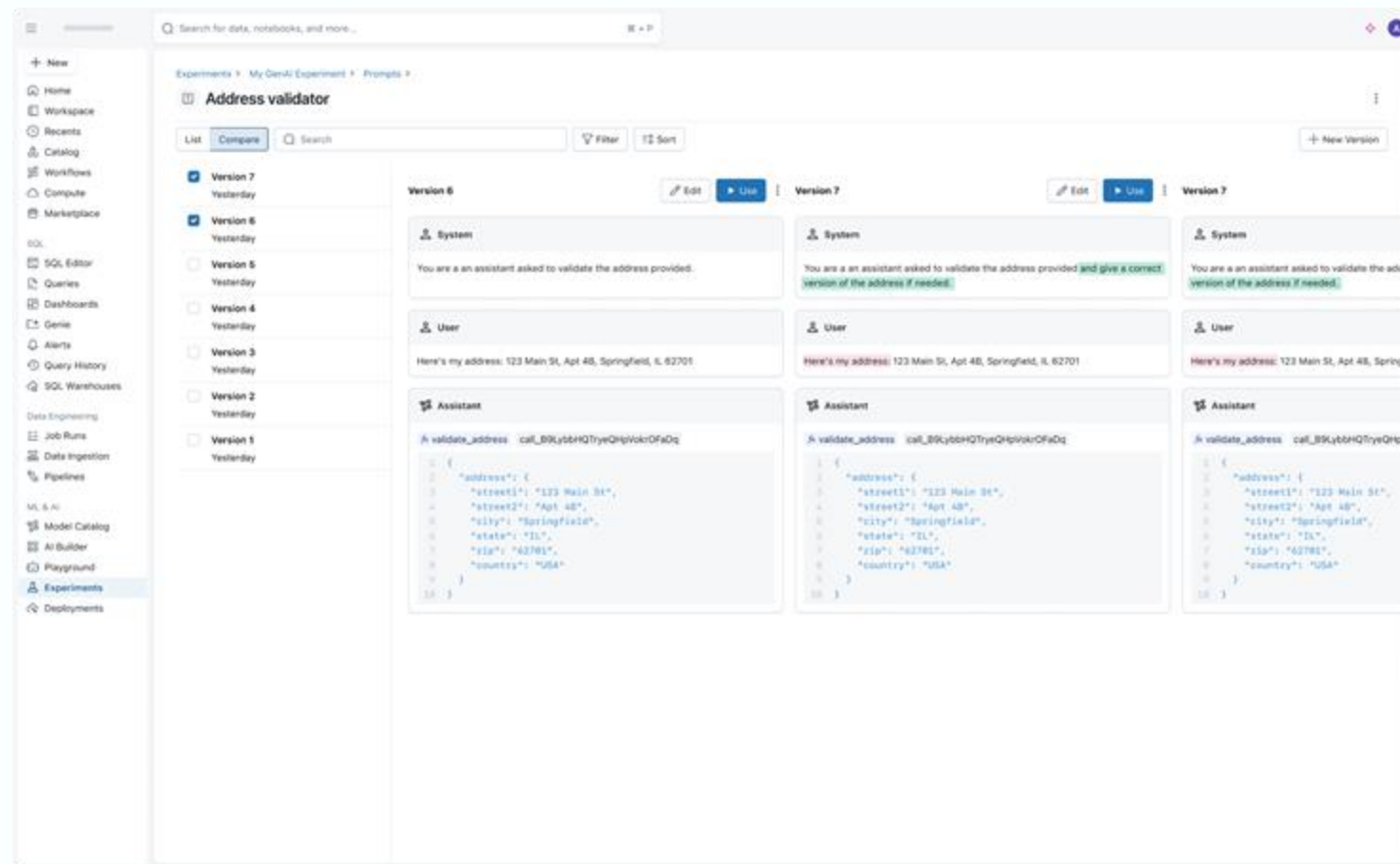
Security govern production traces
with Databricks Unity Catalog



Documentation:
[Databricks](#), [OSS](#)



Prompt Registry



► Prompt management

Version and reuse prompts across agents

```
prompt = mlflow.register_prompt(  
    name="qa",  
    template="Answer the following question: {{question}}",)  
prompt = mlflow.load_prompt(  
    "prompts:/summarization-prompt/2"  
).format(text_to_summarize=query) # Load & format a prompt  
  
response = client.chat.completions.create(  
    messages=[  
        {  
            "role": "user",  
            "content": prompt, # Use the prompt in a chat request  
        }  
    ],  
    model="gpt-4o-mini")
```



► Prompt management with UC

Integrated with Unity Catalog for Enhance governance and discoverability

```
# Register a prompt template
prompt = mlflow.genai.register_prompt(
    name="mycatalog.myschema.customer_support",
    template="You are a helpful assistant. Answer this question: {{question}}",
    commit_message="Initial customer support prompt")

print(f"Created version {prompt.version}") # "Created version 1"

# Set a production alias
mlflow.genai.set_prompt_alias(
    name="mycatalog.myschema.customer_support",
    alias="production",
    version=1
)

# Load and use the prompt in your application

prompt = mlflow.genai.load_prompt(name_or_uri="prompts:/mycatalog.myschema.customer_support@production")
response = llm.invoke(prompt.render(question="How do I reset my password?"))
```



▶ Prompt optimization

Automatically optimize your prompts by leveraging DSPy

```
prompt = mlflow.register_prompt(
    name="qa",
    template="Answer the following question: {{question}}",
)
result = mlflow.genai.optimize_prompt(
    target_llm_params=LLMParams(model_name="openai/gpt-4.1-nano"),
    train_data=[{"inputs": {"question": f"{i}+1"}, "expectations": {"answer": f"{i + 1}"}} for i in range(100)],
    scorers=[exact_match],
    prompt=prompt.uri,
    optimizer_config=OptimizerConfig(num_instruction_candidates=5),
)
```



Experiments >

My GenAI app GenAI

Traces Evaluations Labeling Versions Monitoring

List Compare Search

Filter Sort

+ New version

Select: first 10

Version	Time created	Notebook
<input checked="" type="checkbox"/> Version 20	Mar 24 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 19	Mar 24 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 18	Mar 24 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 17	Mar 24 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 16	Mar 24 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 15	Mar 20 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 14	Mar 20 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 13	Mar 20 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 12	Mar 20 2025 15:58:02 GMT-0700	agent-trial
<input checked="" type="checkbox"/> Version 11	Mar 16 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 10	Mar 16 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 9	Mar 16 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 8	Mar 16 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 7	Mar 16 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 6	Mar 12 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 5	Mar 12 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 4	Mar 12 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 3	Feb 26 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 2	Feb 26 2025 15:58:02 GMT-0700	agent-trial
<input type="checkbox"/> Version 1	Feb 26 2025 15:58:02 GMT-0700	agent-trial

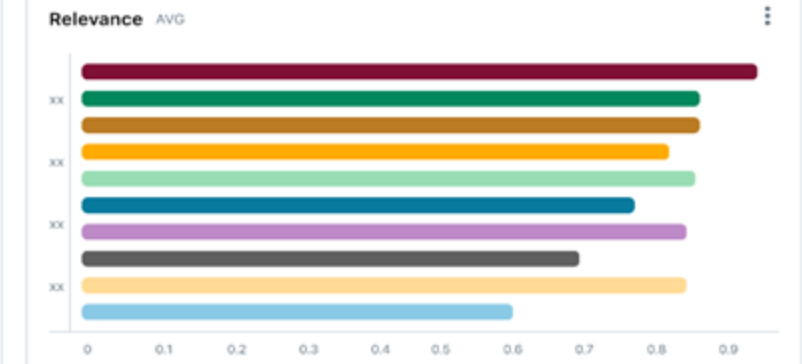
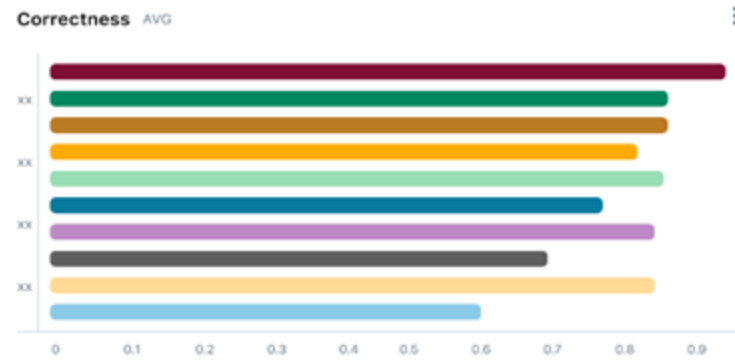
Configuration

Comparison View

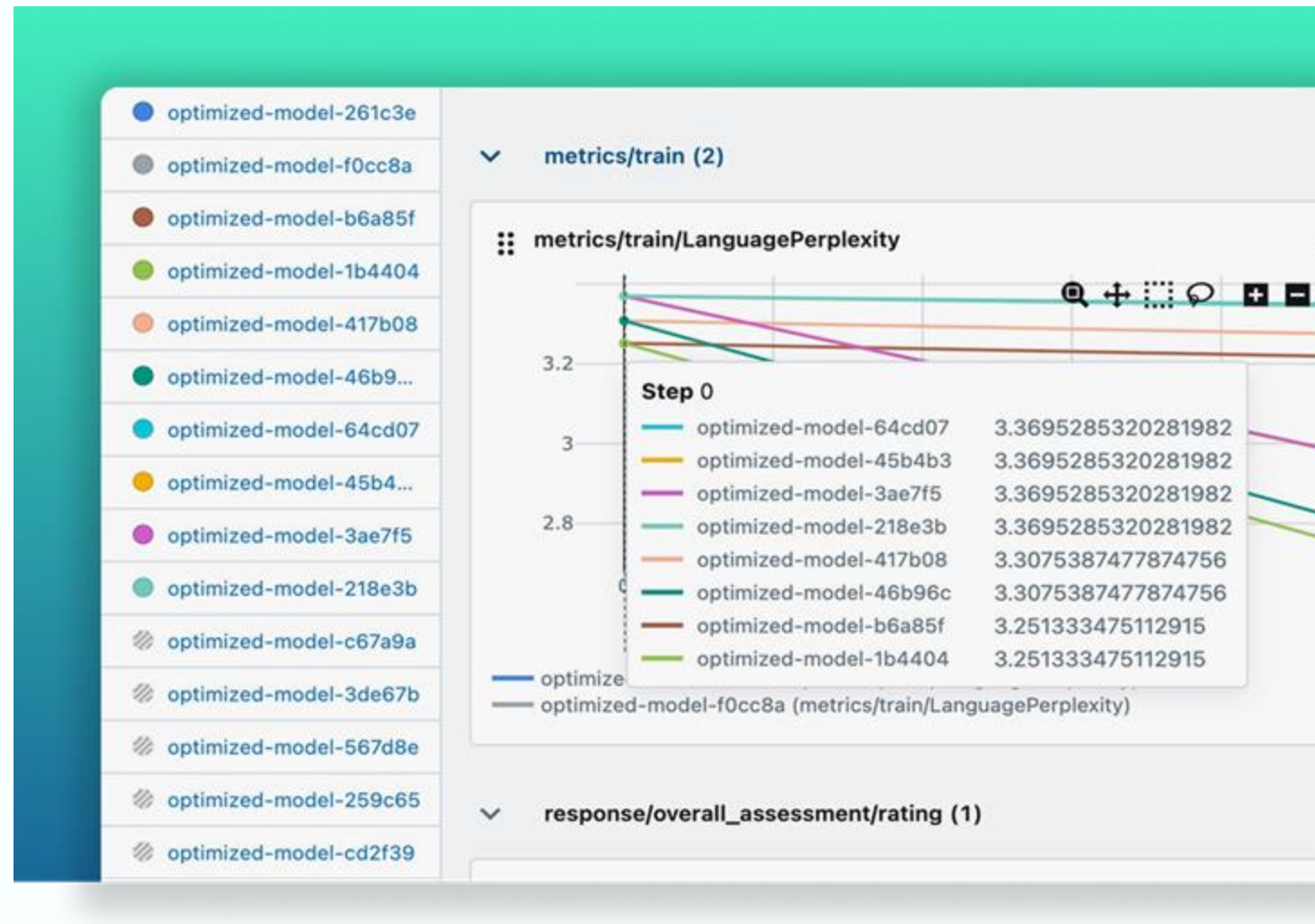
Compare differences

	Version 1 baseline	Version 2	Version 3
Prompt	"I have a question about the company's onboarding process for new employees. Could you provide a detailed overview of the steps involved, including any recent updates to the process and key resources new hires should be aware of? Please pull information from the most recent internal onboarding manuals, training materials, and HR communications."	"I need information about the company's current policies and procedures related to [insert specific topic, e.g., remote work, employee benefits, performance reviews, etc.]. Could you provide a summary of the key points, including any recent changes or updates? Please pull information from the latest internal documents, guidelines, and relevant communications."	"I have a question about the company's onboarding process for new employees. Could you provide a detailed overview of the steps involved, including any recent updates to the process and key resources new hires should be aware of? Please pull information from the most recent internal onboarding manuals, training materials, and HR communications."
Model	GPT-4	GPT-4	GPT-4
Tools	Retrieval Summary	Retrieval Summary	Retrieval Summary

Metric: Dataset 1



ML & DL Improvements



PART 05

行业落地实例

▶ Thousands of our customers have shipped AI in production





replit is an online integrated development environment

Challenge

- They want to use LLMs in their product to assist developers
- Training LLMs are prohibitively expensive and error prone

Solution

- replit built Ghostwriter, a code-generation model, from scratch by training a 2.7B parameter LLM using Pre-training

Impact

- **3 days to train LLM (versus weeks/months)**
- **1 day Raw data to model deployed in production**
- **Lower costs**



Accurate trial data, without manual cleanup

Challenge

AstraZeneca needed a scalable way to **extract insights from 400K+** clinical trial documents to inform competitive and business decisions.

Existing tools were **too slow, too manual, and too costly** to scale.

Solution

They built an AI agent to extract key fields **from 400K+ clinical trial documents**—making insights accessible without relying on engineering.

- Optimized accuracy automatically, **with no retraining**
- Flagged low-quality extractions **using built-in evaluation**

Impact

- **60 minutes** to production
- Enabled non-technical users
- Lowered **cost and complexity**





第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

下一站预告

11/14-15 | 深圳站

12/19-20 | 上海站



查看会议详情

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

AiDD

「深行 · 浅智」

Walk Deep, Think Light.

2025.11.16

AiDD首届麦理浩径徒步





科技生态圈峰会 + 深度研习

——1000+ 技术团队的选择



AiDD峰会详情





第7届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听!

扫码领取会议PPT资料

