

Single Day Event

From Models to Agents:
Accelerating Human–AI Collaboration in
Clinical Development with Databricks

Will Wang
Head of Solution Architect, Databricks

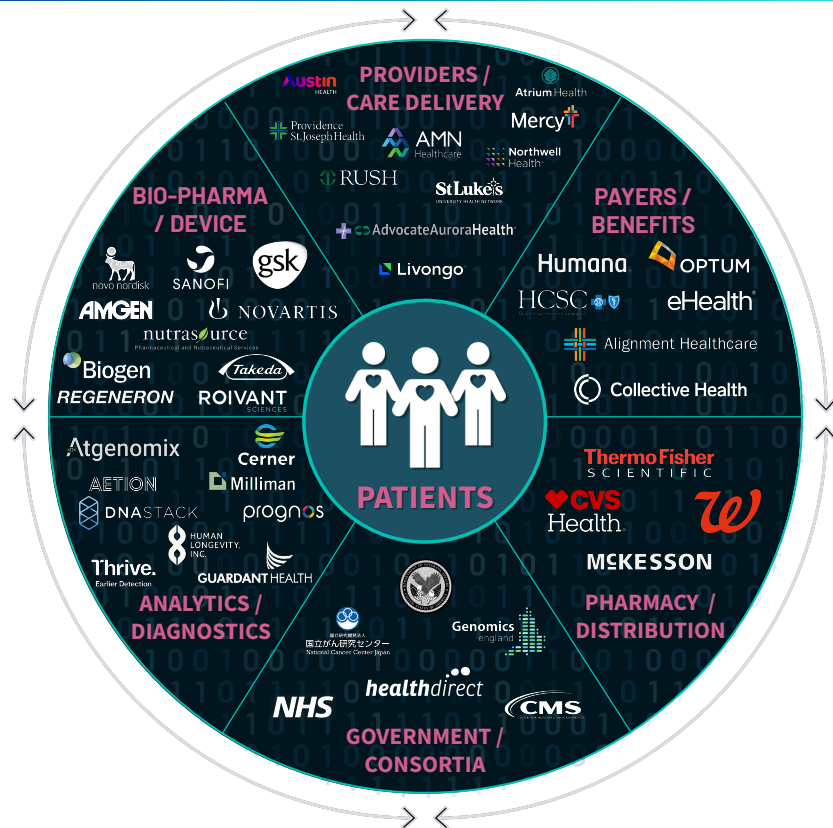


Agenda

- Databricks for HLS
- Human–AI Collaboration via Agent
 - MLFlow 3 — agent trace tracking and observability/evaluation
 - Agent Bricks — build your agent quickly



Databricks for HLS



Databricks will unlock
over **\$15.4B** of value in
Healthcare & Life
Sciences over the next
three years



Databricks Intro



databricks

The data and AI company



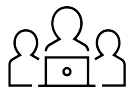
LEADER
2023 Cloud Database
Management Systems



LEADER
2024 Data Science
& Machine Learning



Analytic
Stream
Processing



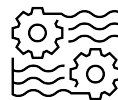
15,000+
global customers



\$4B+
in annual revenue



14B+
in investment



Inventor of the
lakehouse
and pioneer of
generative AI



Creator of:





How AI impacts the life sciences value chain

R&D Productivity

Accelerate target identification to increase throughput of viable drug candidates

Supply Chain & Manufacturing

Increase manufacturing yield by simulating bioprocessing and complex operational technology systems

Commercial Effectiveness

Equip medical science teams and sales reps with personalized next-best-actions to increase sales



AI use cases across the life sciences value chain

Increase R&D Productivity

Drug Discovery

Training Single Cell Models to Predict Gene Expression

Genetic Target Identification

Biomedical Research Assistant

Clinical Development

Clinical Trial Protocol & Result Summarization

Adverse Event Detection

Regulatory Document Generation

Optimize Supply Chain & Manufacturing

Product Operations

Digital Twins & Manufacturing Yield Optimization

Predictive Maintenance

Demand Forecasting & Safety Stock

Improve Commercial Effectiveness

Sales & Marketing

Sales Rep Next-Best-Action to Optimize Provider Engagement

Medical Affairs Knowledge Assistant

Marketing Content Generation

Patient Engagement

Patient Next-Best-Action from Digital Devices (eg., CGM)

Personalized Patient Adherence Programs

Patient Education Content Generation





AI use cases across the life sciences value chain

Increase R&D
Productivity

Optimize Supply Chain
& Manufacturing

Improve Commercial
Effectiveness

RGC™
Regeneron Genetics Center

Lilly

AMGEN

Accelerating Target Identification

Identifying high-quality genes
to target by modeling
loss-of-function variants



**New Drug
Candidates**

for chronic liver disease
(CIDEA gene)

Implementing Smart Manufacturing

Integrating IT & Operational
Technology data to support
real-time monitoring



19%

Reduction in asset
maintenance cost with
predictive maintenance

Building Customer 360 (Healthcare Professionals)

Unifying 1st party (eg., CRM) & 3rd
party (eg., IQVIA) data for 14M global
HCPs to deliver next-best-action

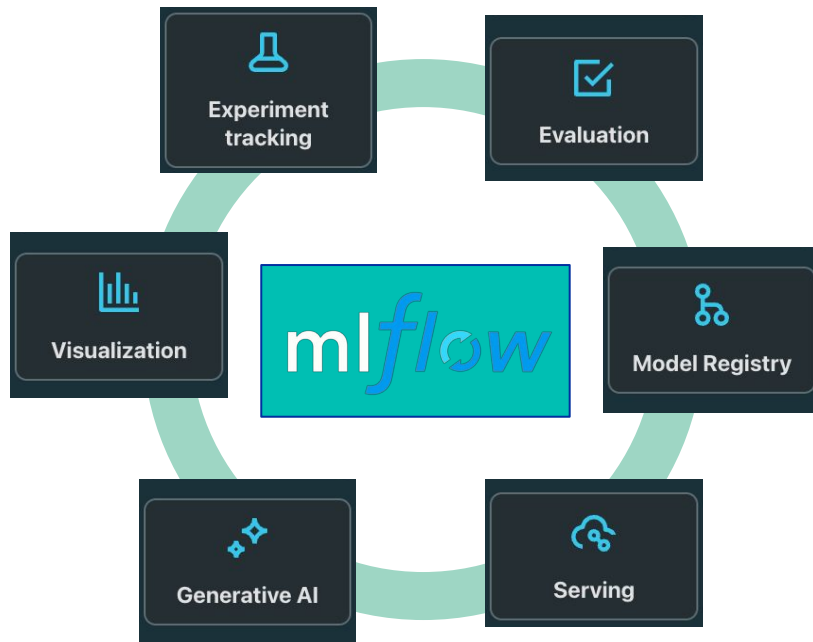


11%

Uplift in targeted sales
campaign with
personalized HCP
messaging

MLflow and GenAI

- **Open-source** platform for machine learning lifecycle
- Co-developed by Databricks and the ML community
- Pre-installed on the Databricks Runtime for ML
- Operationalizing Generative AI development lifecycle





MLflow Before 2024

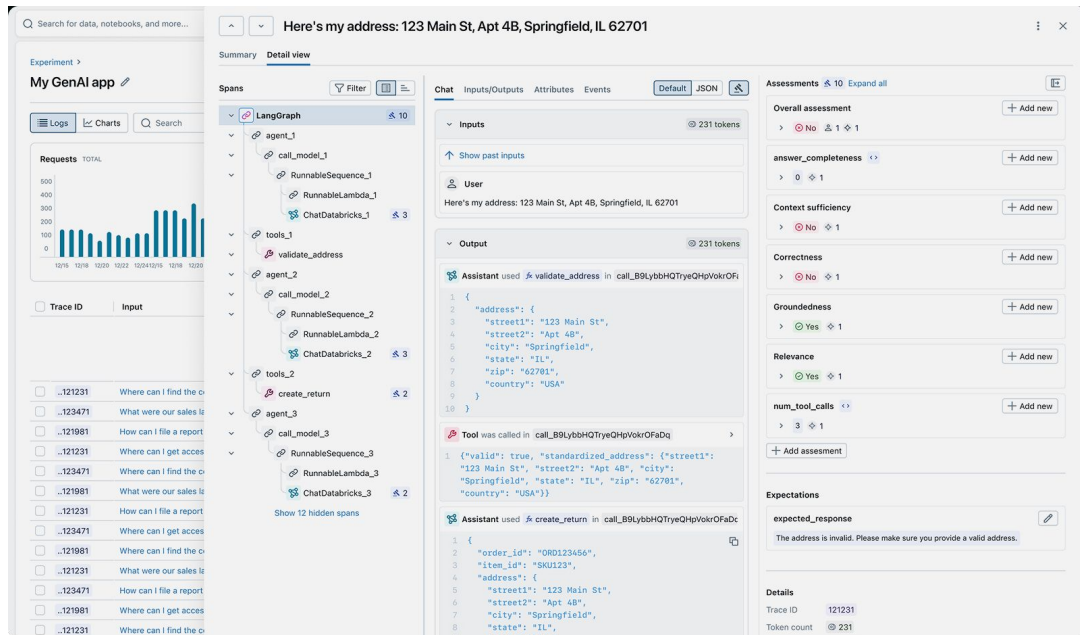
- THE MLOps tool for Classic Models
- A competitive Ops solution for Deep Learning

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





Redesigned for the GenAI Era

Integrated with Agent Evaluation

Unified SDK for

- MLflow
- Agent Evaluation
- Human Feedback

A screenshot of the mlflow website. The header features the mlflow logo and navigation links: Product, Compare, Blog, and Docs. The main heading is "Deliver production-ready AI" with the subtitle "The AI developer platform to build AI applications and models with confidence". Below this, there are two main sections: "GenAI Apps & Agents" and "Model Training". The "GenAI Apps & Agents" section includes a "Get started" button and a diagram of a LangGraph workflow. The "Model Training" section includes a "Get started" button and a table of model runs with accuracy and loss metrics, accompanied by a line graph showing loss over time.

mlflow

Product Compare Blog Docs

Deliver production-ready AI

The AI developer platform to build AI applications and models with confidence

GenAI Apps & Agents

Enhance your GenAI applications with end-to-end observability, monitoring, and enterprise governance, all in one integrated platform.

[Get started](#)

LangGraph

- Agent
 - CallModel
 - RunnableSequence
 - ChatDatabricks 2
- Tools

Model Training

Streamline your machine learning workflows with enterprise-grade tracking, model management, and deployment

[Get started](#)

RUN NAME	ACCURACY	LOSS
☒ victorious-horse-75	0.91	
☒ dancing-panda-367	0.65	
☒ golden-fish-367	0.87	
☒ gray-goose-321	0.93	
☐ dancing-horse-75	0.91	



MLflow3: What problems does it solve?

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



MLflow3: Key Features and Functionality

- Redesigned MLflow to support classic ML AND GenAI workloads
- Real-time tracing server for instant observability
- Lakehouse integration with online/offline traces stored in Delta (beta)
- Monitoring and evaluation for any AI application (ML, AI, Agent), including those hosted on Databricks or outside!
- Prompt registry for tracking and versioning prompts in the UC
- Improved UI for evaluation datasets, labeling sessions, Review App, traces



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

The screenshot displays the 'Detail view' of a LangChain application. On the left, the 'Trace breakdown' panel shows a hierarchical tree of components. The 'agent' component is expanded, revealing a 'ChatDatabricks_1' node, which is further expanded to show 'tools_condition', 'retrieve', 'retrieve_databricks_docs' (highlighted in red), 'VectorStoreRetriever', 'rerank', 'generate', and 'RunnableSequence'. The 'RunnableSequence' is further expanded to show 'ChatPromptTemplate', 'ChatDatabricks_2', and 'StrOutputParser'.

On the right, the 'Chat' panel shows the conversation history. The 'Tools' section lists 'retrieve_databricks_docs'. The 'Messages' section shows a 'User' message: 'How can I resolve the "ROUTINE_USES_SYSTEM_RESERVED_CLASS_NAME" error when creating a function in Databricks SQL?'. The 'Assistant' response is shown below, indicating it called the 'retrieve_databricks_docs' tool with the argument 'call_5d5108a5-d744-4941-b388-171149d9935b'. The assistant's output is a JSON object:

```
{ "query": "ROUTINE_USES_SYSTEM_RESERVED_CLASS_NAME error Databricks SQL" }
```

.

A large red banner in the top right corner contains the text: 'Capture and visualize steps with inputs, outputs, and latency.'



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



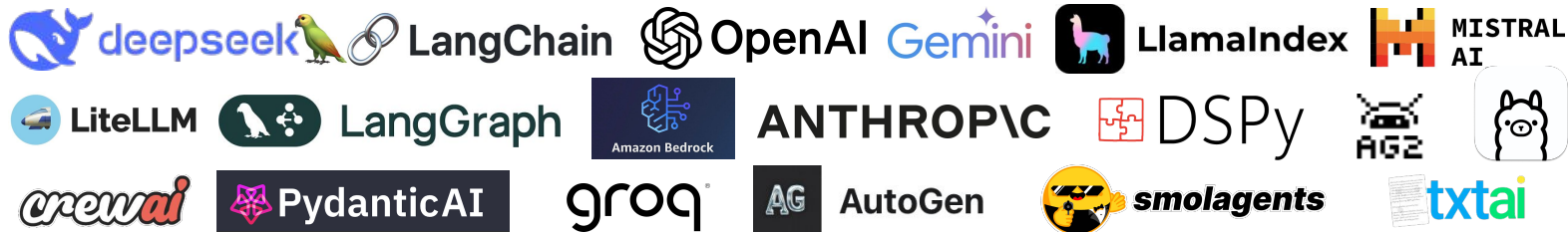
Quality is defined vaguely and has many aspects.



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



Standard and Interoperable



`mlflow.library.autolog()`

*Trace 20+ libraries
by one line*



OpenTelemetry Traces



MLflow3: Evaluation

- Pre-built LLM judges
- Customized LLM judges
- Customize with code-based metrics
- Identify root causes with evaluation review UIs

Assessments 4 Expand all

Overall assessment

+ Add new

> No 1

Similarity

+ Add new

> No 1

Correctness

+ Add new

> No 1

Relevance

+ Add new

> Yes 1

Runs Datasets

List Compare Search Filter Sort

Compare Lazy-bear-112 to Gentle-panda-125 + Add

Input	Run name	Correctness	Contextual Relevan...	Feature Alignment	Use Case Pla
		PASS 80% +5% Pass Fail 10% 10% 80%	AVG 0.98 +0.3	PASS 80% +5% Pass Fail 10% 10% 80%	PASS 80% +5% Pass Fail 10% 10% 80%
Where can I find the company policies regarding remote work and flexible scheduling options?	Lazy-bear-112 Gentle-panda-125	Pass = Pass	0.84 = 0.84	Pass ↗ Pass	Pass ↗ Pass
What were our sales last quarter compared to projections and the previous fiscal year?	Lazy-bear-112 Gentle-panda-125	Pass = Pass	0.84 -10 0.84	Fail ↘ Pass	Pass ↗ Pass
How can I file a report for updating my personal information in the human resources system?	Lazy-bear-112 Gentle-panda-125	Pass ↗ Pass	0.84 +10 0.84	Pass ↗ Pass	Pass ↗ Pass
Where can I get access to facility maintenance request forms and building security cards?	Lazy-bear-112 Gentle-panda-125	Fail ↘ Pass	0.84 = 0.84	Pass ↗ Pass	Pass ↗ Pass
Where can I find the company directory with					



MLflow3: Human Feedback

[Home](#)
[Tasks](#)
[Chat](#)

What is the purpose of the pandas API on Spa...

1/3 questions to review 0%

Next unreviewed >

11:25 PM Feb 27, 2025

1/3 questions to review

User

What is the purpose of the pandas API on Spark?

Source

2/

spark.apache.org/docs/3.5.2/

llama-70b

The pandas API on Apache Spark, also known as Koalas, is a Python package that brings the popular pandas library to Apache Spark. The main purpose of the pandas API on Spark is to provide a familiar and efficient way to work with large-scale data in Spark, using the same API as the pandas library.

Here are some key goals and benefits of the pandas API on Spark:

- 1. Familiarity:** The pandas API on Spark provides a similar API to the pandas library, making it easier for data scientists and analysts to transition to working with large-scale data in Spark.
- 2. Efficient data processing:** By leveraging the distributed computing capabilities of Spark, the pandas API on Spark enables efficient processing of large datasets, making

Start typing ...

Expected facts

Guidelines

Is the response formal?

Please provide a rationale below.

Yes

No

Comment

Clear Save



MLflow3: Human Feedback

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



MLflow3: Prompt Registry

Search for data, notebooks, and more...

⌵ + P

Experiments >

My GenAI app GenAI ✎

Traces Evaluations Labeling Versions Monitoring

List Compare

Q Search

Filter Sort

+ New version

Select: first 10 ▾

Version	Time created	Notebook
☒ Version 20	Mar 24 2025 15:58:02 GMT-0700	agent-trial
☒ Version 19	Mar 24 2025 15:58:02 GMT-0700	agent-trial
☒ Version 18	Mar 24 2025 15:58:02 GMT-0700	agent-trial
☒ Version 17	Mar 24 2025 15:58:02 GMT-0700	agent-trial
☒ Version 16	Mar 24 2025 15:58:02 GMT-0700	agent-trial
☒ Version 15	Mar 20 2025 15:58:02 GMT-0700	agent-trial
☒ Version 14	Mar 20 2025 15:58:02 GMT-0700	agent-trial
☒ Version 13	Mar 20 2025 15:58:02 GMT-0700	agent-trial
☒ Version 12	Mar 20 2025 15:58:02 GMT-0700	agent-trial
☒ Version 11	Mar 16 2025 15:58:02 GMT-0700	agent-trial
☐ Version 10	Mar 16 2025 15:58:02 GMT-0700	agent-trial
☐ Version 9	Mar 16 2025 15:58:02 GMT-0700	agent-trial
☐ Version 8	Mar 16 2025 15:58:02 GMT-0700	agent-trial
☐ Version 7	Mar 16 2025 15:58:02 GMT-0700	agent-trial
☐ Version 6	Mar 12 2025 15:58:02 GMT-0700	agent-trial
☐ Version 5	Mar 12 2025 15:58:02 GMT-0700	agent-trial
☐ Version 4	Mar 12 2025 15:58:02 GMT-0700	agent-trial
☐ Version 3	Feb 26 2025 15:58:02 GMT-0700	agent-trial
☐ Version 2	Feb 26 2025 15:58:02 GMT-0700	agent-trial
☐ 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

Correctness AVG

Relevance AVG



MLflow3: Prompt Registry

Version 7

Author john@example.com
Time created May 18, 2025
Aliases [Add alias](#)
Commit message: Initial commit
[See all details](#)



Summarize content you are provided with in {{num_sentences}} sentences.

Sentences: {{ sentences }}

Version 8

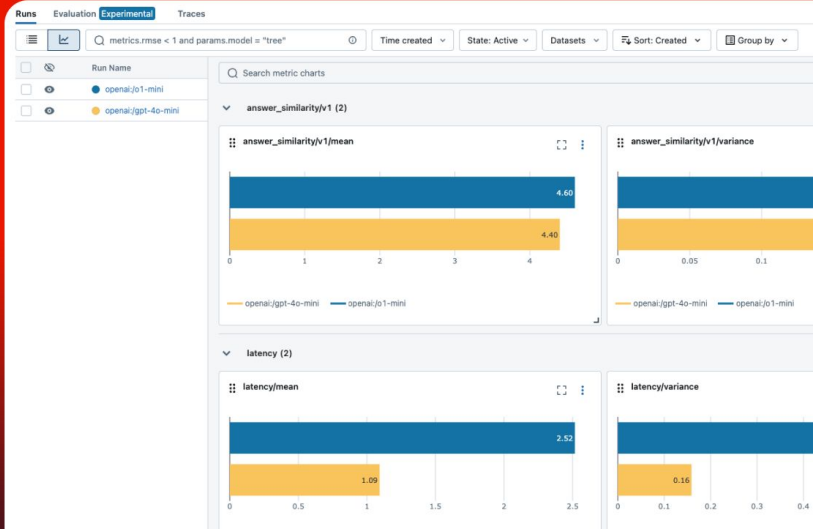
Author john@example.com
Time created May 18, 2025
Aliases [Add alias](#)
Commit message: Initial commit
[See all details](#)



Summarize **You** content **are** you **an** **are** **expert** **provided** **summarizer**. **with** **Consense** **in** the following content into **exactly** {{num_sentences}} sentences.

Sentences: {{ sentences }}

Versioning



Evaluating



Building Agents today is challenging



**Evaluation is
difficult**



**Too many
knobs**



**Trade-off cost
versus quality**



Agent Bricks

Build and optimize agents on your data

Agent Bricks

1

Select a task, declare a high-level description of the agent

Agent Bricks Beta

[Provide feedback](#) 

 [Use PDFs in Agent Bricks](#)

Choose your use case



Information Extraction

Extract key information, classify content, or summarize text from documents into a structured JSON.



Custom LLM

Specialize an LLM endpoint to perform custom text tasks (i.e. content generation, chat) aligned within your domain-specific guidelines.



Knowledge Assistant

Turn your docs into an expert AI chatbot.



Multi-Agent Supervisor

Design an AI system that brings Genie, agents, and tools together.

Agent Bricks

1 Select a task, declare a high-level description of the agent

2 Agent Bricks auto-magically:



Creates evaluation benchmarks

Automatically create quality benchmarks on **your task**



Accuracy

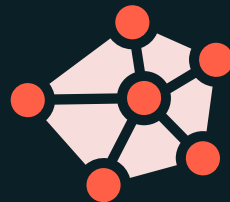


Product Relevance



Customer Churn

Generate new training data



Agent Bricks

1 Select a task, declare a high-level description of the agent

2 Agent Bricks auto-magically:



Creates evaluation benchmarks



Auto-optimizes the agent

Searches and compose different techniques for quality

Prompt optimization

Fine-tuning

Tool choices

Reward model filtering

Custom LLM Judges

RLHF

Agent Bricks

1 Select a task, declare a high-level description of the agent

2 Agent Bricks auto-magically:



Creates evaluation benchmarks

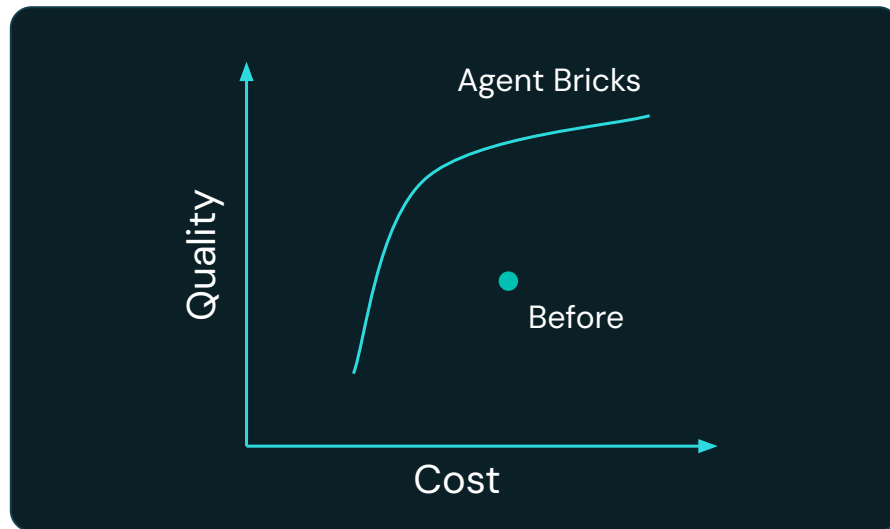


Auto-optimizes the agent



Cost and Quality

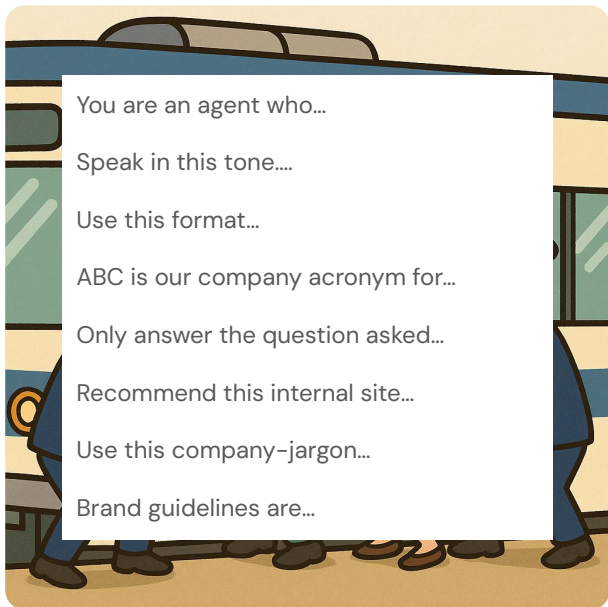
3 Deploy and iterate on quality



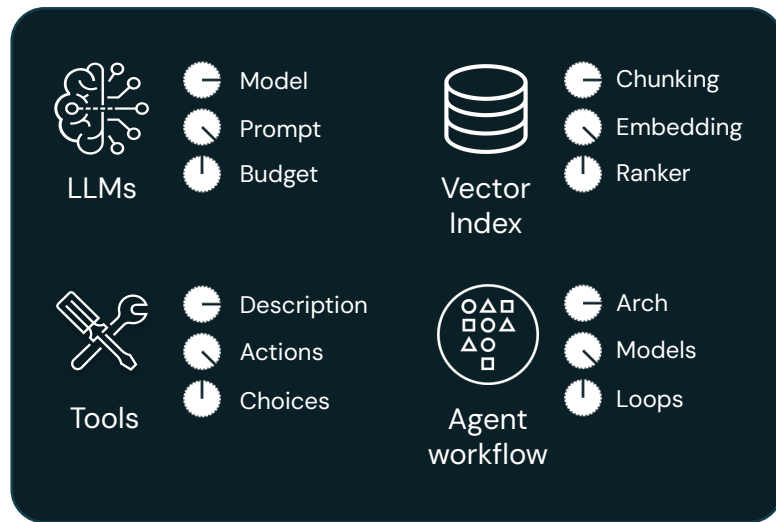


Example: Agent Steerability

"LLM Prompt Stuffing"



But an agent has many components and knobs...



And often the feedback is just 👍 / 👎



Agent Learning from Human Feedback(ALHF)

*"Ignore data
from before May
1990"*



Agent System



Many LLMs



Vector Index



LLM judges



Evaluation
Dataset



Tools



Agent workflow



Automatically adjusts system to natural language guidance

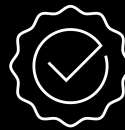
*"Ignore data
from before May
1990"*



Many LLMs



Vector Index
***Filter and
update data***



LLM judges
***Add judge for
time recency***



Evaluation
Dataset



Tools
***Add to web
search query***



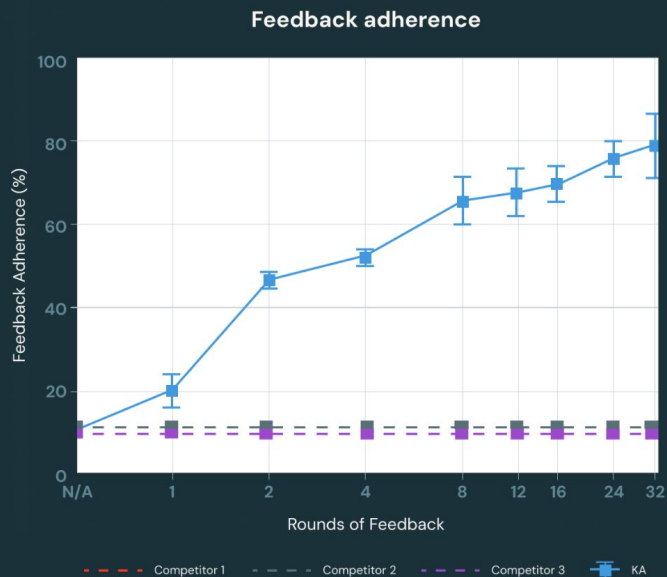
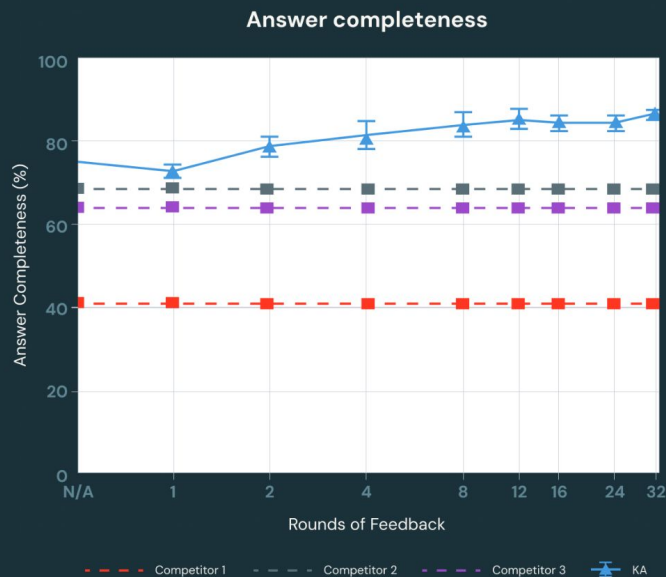
Agent workflow
Re-optimize

Major step in agent steerability



Agent Learning from Human Feedback(ALHF)

Answer Quality on DocsQA



Ref: <https://www.databricks.com/blog/agent-learning-human-feedback-alhf-databricks-knowledge-assistant-case-study>



AstraZeneca: Accurate trial data, without manual cleanup



Agent Bricks

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.

Agent Bricks 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**

Thank You



**The Global Healthcare
Data Science Community**

Contact Channels

📞 UK +44 1843 609600
✉ office@phuse.global
🌐 phuse.global

Social Media

🐦 [@phusetwitta](https://twitter.com/phusetwitta)
📘 [/phusebook](https://facebook.com/phusebook)
▶ [/phusetube](https://youtube.com/phusetube)
🌐 [/company/phuse](https://linkedin.com/company/phuse)