fiddler

Guide

The Ultimate Guide to LLM Monitoring

Ensure high performance, behavior, and safety of LLM applications

Fiddler AI Observability Platform for LLMOps

Fiddler is the pioneer in enterprise AI Observability and offers a comprehensive LLMOps platform that aligns teams across the organization to deliver high performing and responsible models and applications. The Fiddler AI Observability platform helps developers, platform engineering, and data science teams through the lifecycle to evaluate, monitor, analyze, and protect models and applications.

Fiddler helps organizations harness the power of generative AI to deliver correct, safe, and secure chatbots and LLM applications to:

- Increase process automation
- Support customer service and engagement
- Enhance employee decision making and experience, and more

Fortune 500 organizations use Fiddler to deliver high performance AI, reduce costs and increase ROI, and be responsible with governance.

Key Enterprise Concerns on AI

Enterprises are leveraging generative AI and LLMs to grow their business, maximize revenue opportunities, automate processes, and improve customer and employee satisfaction.

As these enterprises launch LLM-based applications, they also need to address concerns surrounding generative AI like performance, quality, safety, privacy, correctness and among others. By addressing these concerns prior to launching LLM applications, developers, platform engineering and business teams can deliver performant, helpful, safe, and secure LLMs to end-users while derisking adverse outcomes.

Performance
How satisfactory is the model's response?

Quality
How is the data quality?

Safety
Are the user prompt and model responses safe?

Cost
Where is the best ROI?

Correctness
How accurate is the response?

Transparency
Why did the model say that?

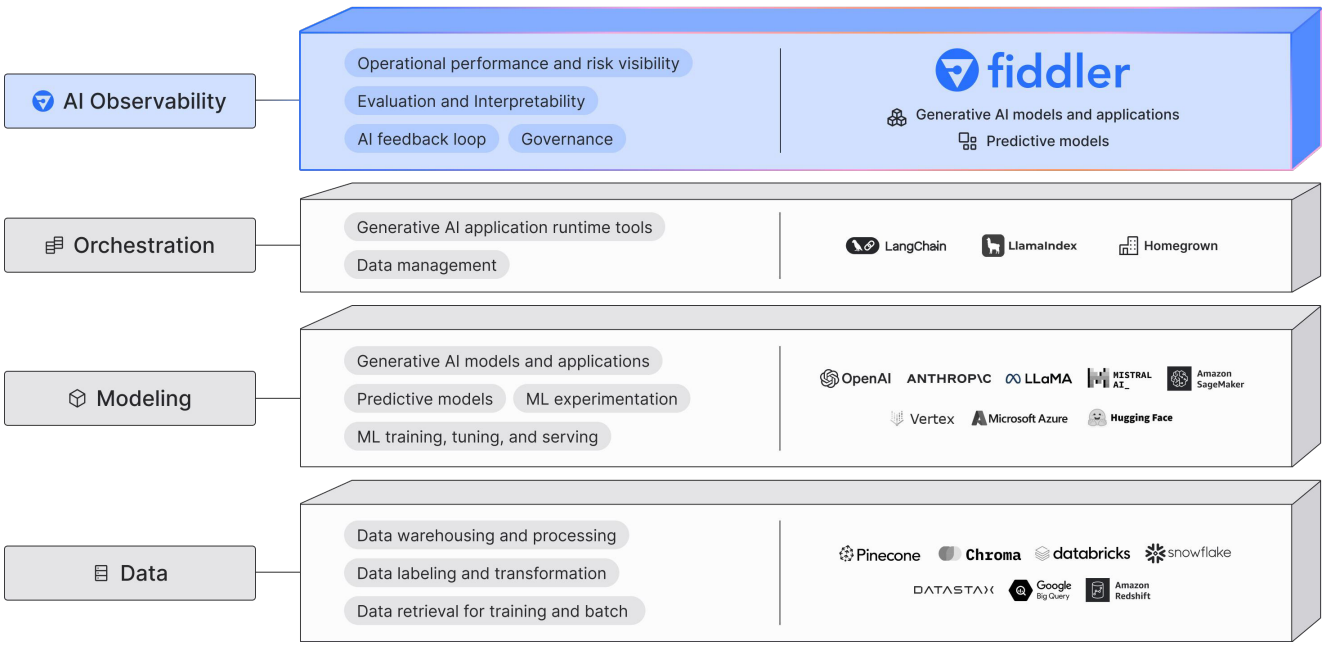
Bias
Is the model's response biased?

A/B Test
Is the model changing across versions?

Privacy
Is the model leaking private data?

Robustness
How sensitive is the model's response?

The New MOOD Stack for LLMOps



The MOOD stack is the new stack for LLMOps to standardize and accelerate LLM application development, deployment, and management. The stack comprises Modeling, AI Observability, Orchestration, and Data layers that are essential for LLM powered applications. Enterprises adopting the MOOD stack for scaling their deployments gain improved efficiency, flexibility, and enhanced support.

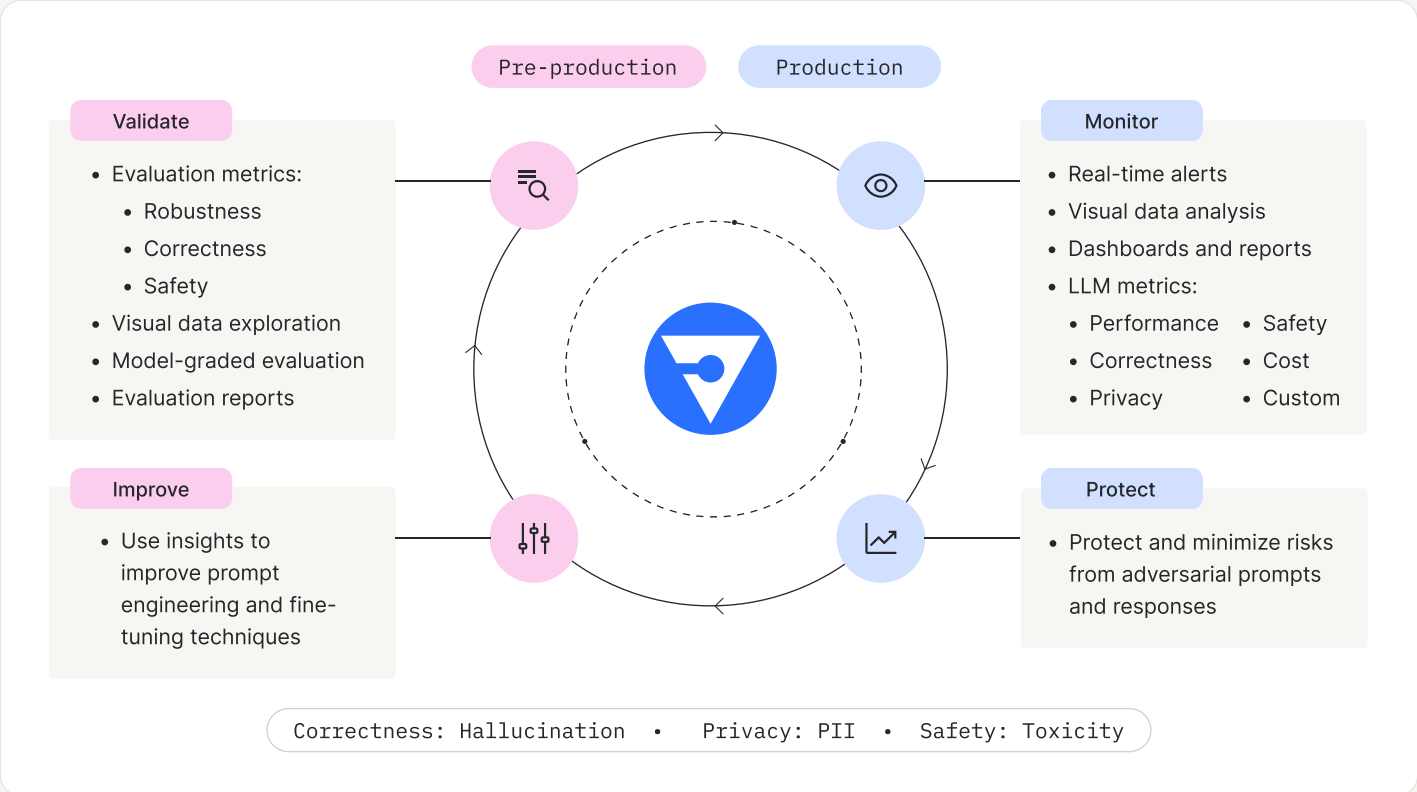
AI Observability is the most critical layer of the MOOD stack, enabling governance, interpretability, and the monitoring of operational performance and risks of LLMs. This layer provides the visibility and confidence for stakeholders across the enterprise to ensure production LLMs are performant, safe, correct, and trustworthy.

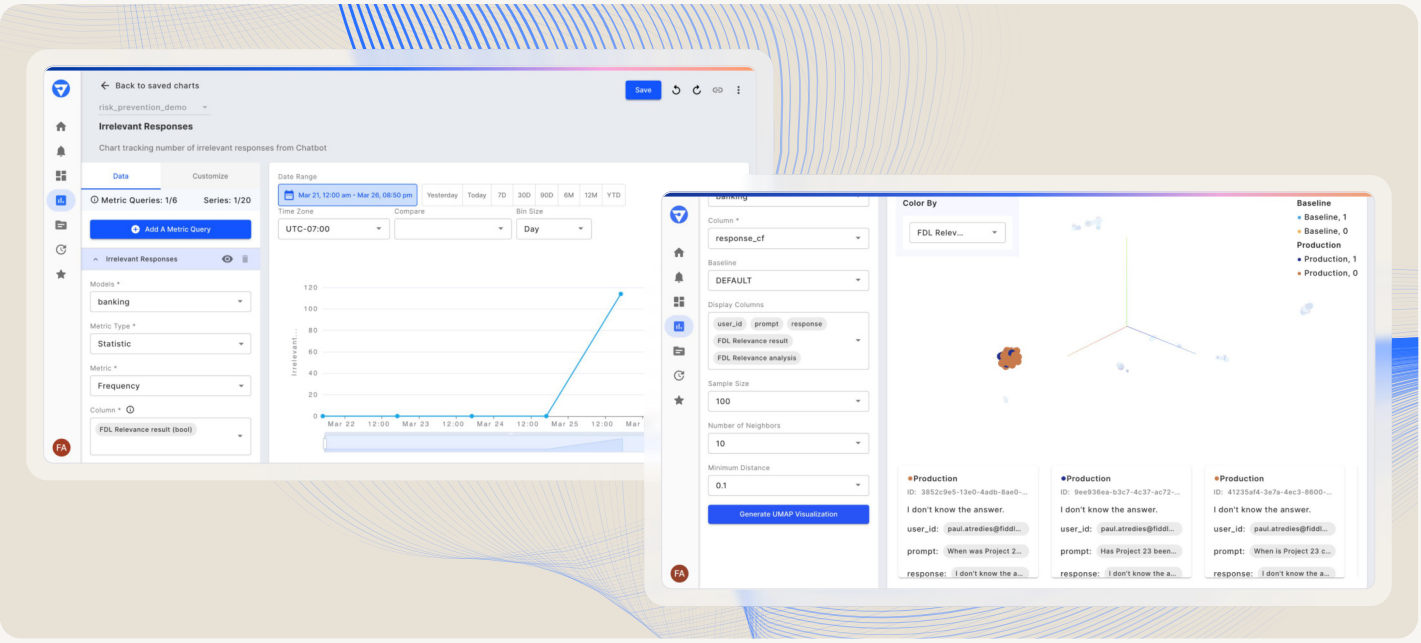
The AI Observability layer is the culmination of the MOOD stack, enhancing enterprises' ability to maximize the value from their LLM deployments.

Comprehensive AI Observability Platform for LLMOps

The Fiddler AI Observability platform is designed and built to help customers address the concerns surrounding generative AI.

Whether AI teams are launching AI applications using open source, in-house built LLMs or commercial LLMs, Fiddler equips users across the organization with an end-to-end LLMOps experience, spanning from pre-production to production. With Fiddler, you can evaluate, monitor, analyze, and protect large language models and applications.





Fiddler offers a comprehensive, enterprise-grade AI Observability platform to help organizations build the foundation for an end-to-end LLMOps. Monitor, analyze, and protect LLMs in production. Detect and resolve issues, like hallucinations, adversarial attacks, and data leakage, to minimize risks impacting users from adversarial model outcomes.

Key Capabilities

Actionable Alerts

Improve operational efficiency with alerts and automated monitoring

Drift Monitoring

Detect changes in LLM performance and behavior caused by data drift

LLM Metrics

Measure metrics, such as hallucination, groundedness, faithfulness, feedback, toxicity, PII, sentiment, profanity, cost, and custom

3D UMAP Analysis

Visualize qualitative insights by identifying data patterns and trends on a 3D UMAP, including human-in-the-loop direct and indirect user feedback

Root Cause Analysis

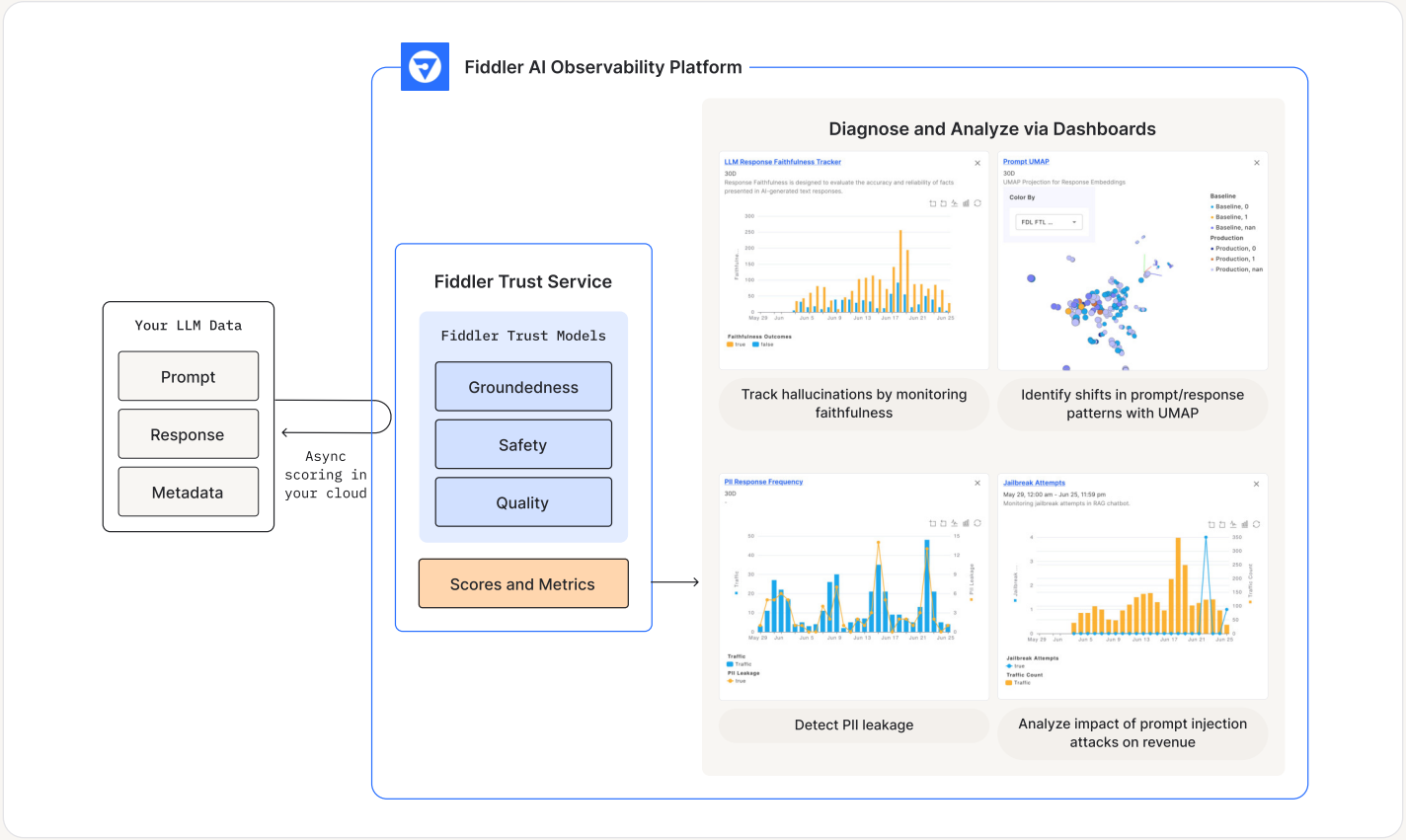
Perform root cause analysis to obtain a complete dataset of problematic prompts/responses and see how they impacted model performance

Dashboards and Charts

Create dashboards and reports that track PII, toxicity, hallucination, and other LLM metrics to increase cross-team collaboration to improve LLMs

Fiddler Trust Service for LLM Metrics Monitoring

Fiddler offers a comprehensive library of LLM metrics, or a LLM trust service, to measure and surface issues in prompts and responses. Model developers and application engineers can customize their monitoring by using Fiddler Trust Services and selecting specific LLM metrics tailored to their use cases. Under the hood, Fiddler Trust Models, proprietary fine-tuned models, evaluate inferences from the LLM application and provide a score of both prompts and responses based on the chosen LLM metrics, ensuring comprehensive metrics monitoring.



These LLM metrics can also be plotted on a 3D UMAP visualization with embedding generation to track prompt and response outliers and drift:

Hallucination Metrics

- Faithfulness/ Truthfulness/ Groundedness
- Answer relevance
- Context relevance
- Conciseness
- Coherence

Safety Metrics

- PII
- Toxicity
- Jailbreak
- Sentiment
- Profanity
- Regex match
- Topic
- Banned keywords
- Language detection

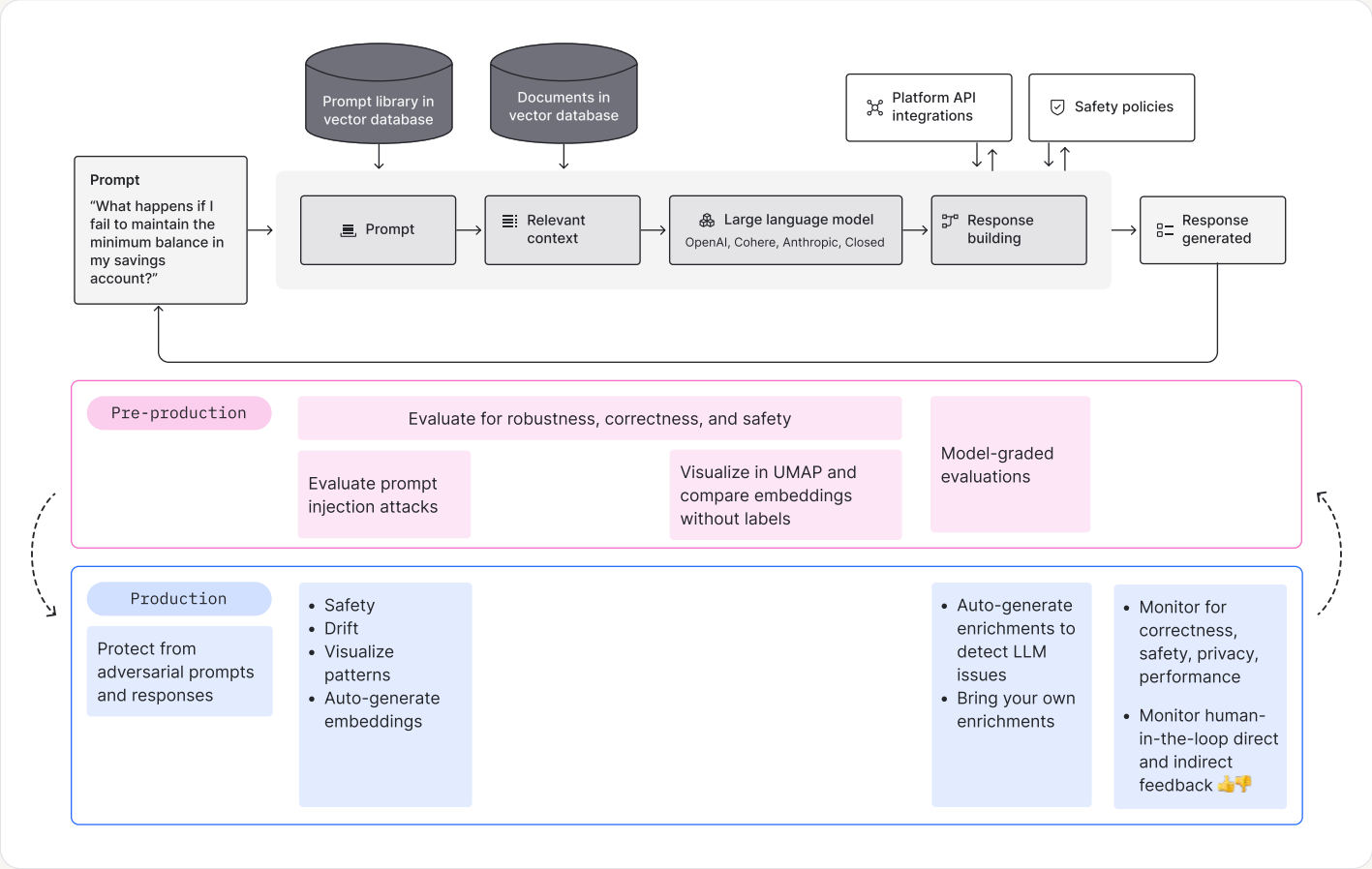
Operational Metrics

- Cost (tokens)
- Latency
- Session length

How Fiddler Works in the RAG Architecture

Depending on the AI strategy and use case, there are four ways organizations deploy LLMs, including prompt engineering with context, retrieval augmented generation (RAG), fine-tuning, and training. RAG is a common approach to deploy an LLM application as it's effective in improving the quality of responses generated by an LLM.

Fiddler helps organizations launch LLM-powered chatbots and applications throughout the LLMOps lifecycle, from pre-production to production, regardless of what LLM deployment method they use.



LLM Trust Standards for Enterprises

Enterprises deploying LLMs must rigorously adhere to the six LLM trust standards to ensure secure, ethical, and compliant AI operations. These standards are essential for safeguarding data privacy and enhancing the reliability of AI applications.

Secure Data Retrieval and Dynamic Grounding

Ensure that data access via generative AI is restricted to authorized personnel to safeguard information confidentiality. Implement stringent verification of information sources to ensure accuracy and relevance of AI-generated content.

Data Masking

Employ strategies to anonymize personal information within AI processes, thus protecting privacy and maintaining corporate integrity.

Prompt Defense

Set up prompt defense guardrails to protect against bad actors and harmful outputs, including prompt instructions.

Toxicity Detection

Evaluate AI outputs to identify and mitigate offensive or harmful content prior to deployment.

Zero-Data Retention

Establish robust security controls and zero-data retention policy agreements to ensure prompts and responses are erased and never retained.

Audit Trail

Maintain an audit trail to review AI usage and development practices to assure compliance with evolving legal and ethical standards.

Your Partner for AI Observability for LLMOps

One Unified Platform

- A full stack, actionable AI Observability platform that supports predictive and generative models
- Designed for Engineering, Data Science, LOB, Risk & Compliance, and Ethics teams

Comprehensive AI Observability

- Supports end-to-end AI Observability from pre-production to production to reporting
- Supports visual exploratory data analysis (EDA) on unstructured data

LLM Metrics for Model Health

- Measure LLM health and performance with LLM metrics like safety, privacy, and correctness
- Monitor custom LLM metrics to meet business KPIs

Built for the Enterprise

- Enterprise-grade scalability and stability
- SaaS and VPC
- Customizable dashboards and reports for cross-team collaboration and decision-making

Expert AI Team

- Design partnerships and build responsible AI practices
- White glove support for advanced AI strategy and applications

fiddler

Fiddler is a pioneer in AI Observability for responsible AI. The unified environment provides a common language, centralized controls, and actionable insights to operationalize ML/AI with trust. Monitoring, explainable AI, analytics, and fairness capabilities address the unique challenges of building in-house stable and secure LLM and MLOps at scale.

Fiddler helps you grow into advanced capabilities over time and build a framework for responsible AI practices.

Fortune 500 organizations use Fiddler across pre-production and production to deliver high performance AI, reduce costs, and be responsible in governance.

fiddler.ai sales@fiddler.ai