

Fiddler Trust Service for AI Observability and Security

Purpose-built trust models for secure, in-environment evaluation powering agentic and LLM observability, and guardrails.








CSAA Insurance Group,
a AAA Insurer




AI Observability and Security is the foundation which gives enterprises the confidence to ship more AI agents, and generative and predictive applications into production safely and responsibly.

Integral to the platform, the Fiddler Trust Service runs a series of purpose-built, “batteries-included” Fiddler Trust Models at runtime to evaluate agent and LLM inputs and outputs, enabling tests and experiments in pre-production, monitoring in production, and the industry’s fastest guardrails. The Fiddler Trust Service can be deployed in cloud or VPC environments to maintain strict data control and safeguard agents and LLM applications. Because out-of-the-box and custom enrichments stay within your environment, there’s no risk of data exposure and no Trust Tax from external API calls.

Fiddler Trust Models Deliver

<100ms
Guardrails
Response Time

7-18x
Cheaper*

5M+
Requests
Per Day

Enterprise Secure

Zero data egress from the customer’s environment

Why Enterprise Leaders Choose the Fiddler Trust Service for Agentic and LLM Observability, and Guardrails

Fastest

With a <100ms latency at runtime, the models are optimized for rapid scoring, monitoring, and guardrails, ensuring enterprises can quickly detect and resolve agentic and LLM issues.

Secure

Fiddler can be deployed in cloud or VPC environments, maintaining compliance and protecting sensitive data.

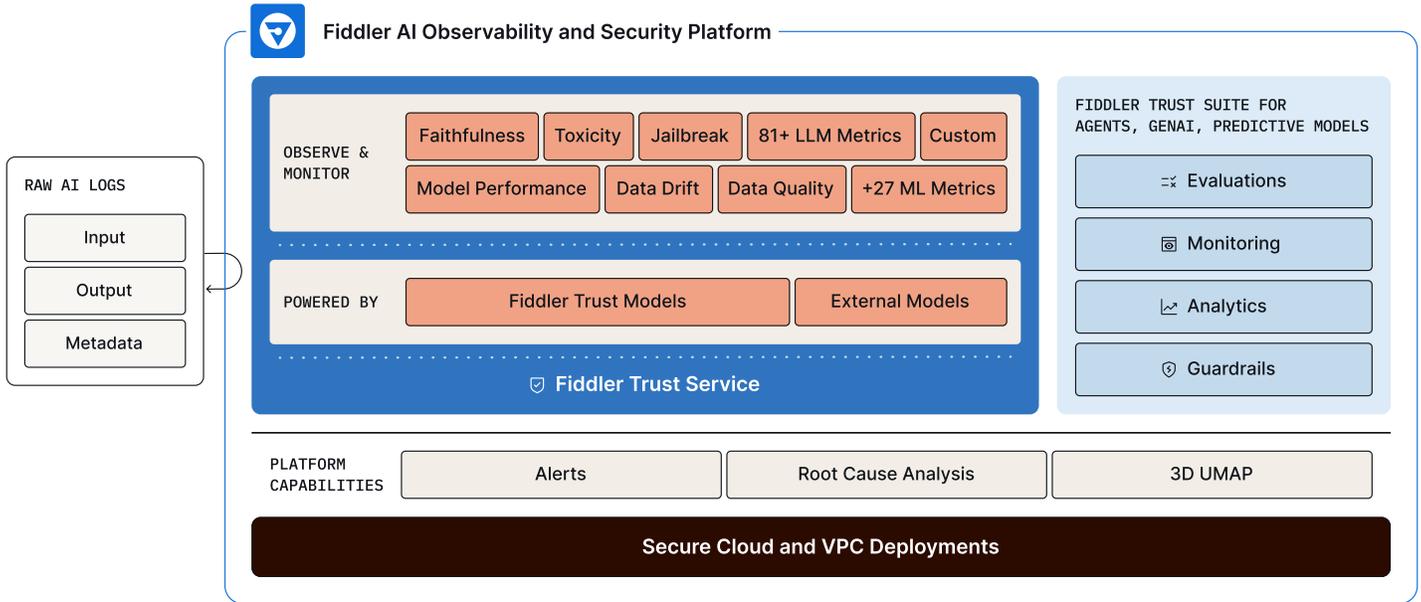
Cost-Effective

Trust Models eliminate the Trust Tax from external API calls while optimizing for efficiency, accuracy, and minimal computational overhead.

Scalable

Built to support high-volume enterprise workloads, Fiddler Trust Service handles over 5 million requests per day out-of-the-box, enabling seamless enterprise deployments at scale.

*Fiddler Trust Models are benchmarked against publicly available datasets.



Fiddler Trust Service: Quality and Moderation Controls for Agents and LLM Applications

The Fiddler Trust Service is an enterprise-grade solution that enables efficient use of computational resources and helps control costs compared to other LLM-as-a-judge offerings.

It combines two main scoring approaches: Fiddler Trust Models and external judges. Together, they deliver over 80 LLM metrics out-of-the box and custom metrics, powering agentic observability, LLM observability, and guardrails for accurate, high-performing, and secure enterprise deployments.

MODERATION CENTER

Guardrails

How does Fiddler score LLM?

Fiddler scores LLMs using the Fiddler Trust Models, which are proprietary fine-tuned models designed to quickly calculate scores for user prompts and LLM responses. These Trust Models are optimized for speed, safety, cost, and task-specific accuracy, delivering near real-time calculations. Unlike closed source LLMs with hundreds of millions of parameters that are highly expressive but not optimized for specific tasks, the Fiddler Trust Models are efficient and effective in scoring specific metrics. This efficiency allows enterprises to scale their GenAI and LLM deployments across their organization effectively.

The Fiddler Trust Models generate enrichments to calculate prompts and responses, enhancing user prompts and LLM responses with supplementary information for accurate scoring. This process involves utilizing the Trust Models to generate scores that help in evaluating the performance, correctness, and safety of the LLMs. By leveraging these models, enterprises can create a continuous feedback loop to improve their LLMs and ensure better outcomes.

Answer Faithfulness: 0.998 Jailbreak Likelihood: 0.001 Guardrail Latency: 99.6 ms

- Deploy the fastest guardrails securely at <100ms.
- Out-of-the-box integrations with NVIDIA NeMo.

QUALITY CENTER

LLM Observability

Jailbreak Tracker

Track total number of jailbreak attempts detected and blocked.

Traffic: Jailbreak Attempts, True

- Monitor hallucination, safety, jailbreak attempts, PII/PHI, and 80+ metrics.
- Get real-time alerts on LLM issues.

QUALITY CENTER

Agentic Observability

- Gain deep visibility into AI agent behaviors.
- Interactive span-level drill-down analytics.

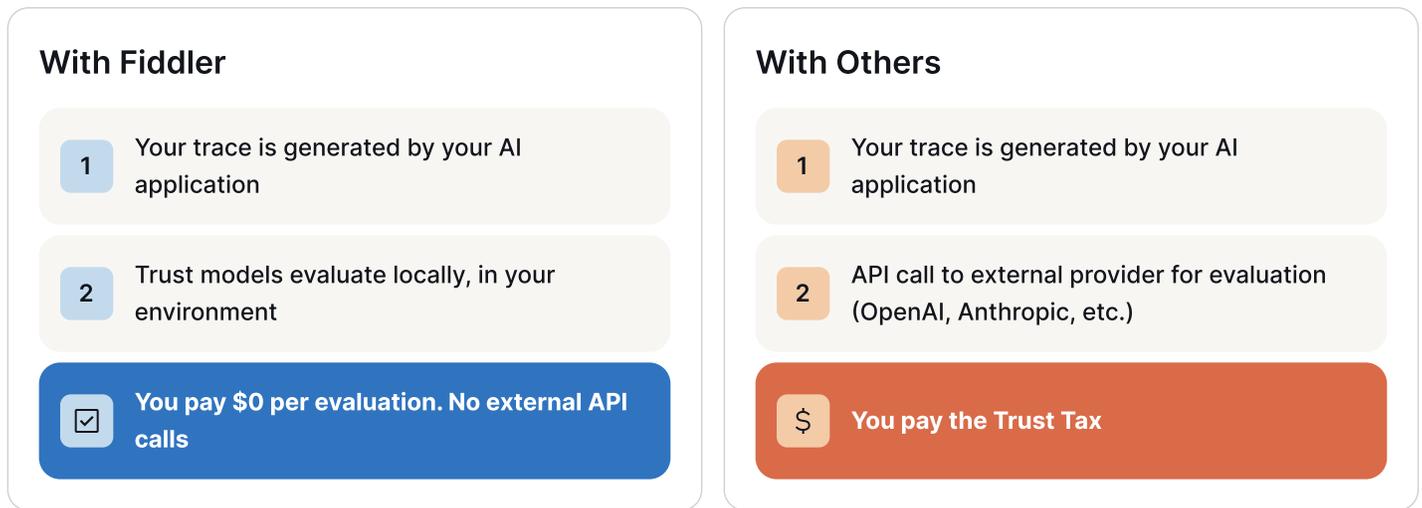
Fiddler Trust Models

In-house built, "batteries-included" trust models for fast, cost-effective, accurate and secure evaluation deployed in your environment.

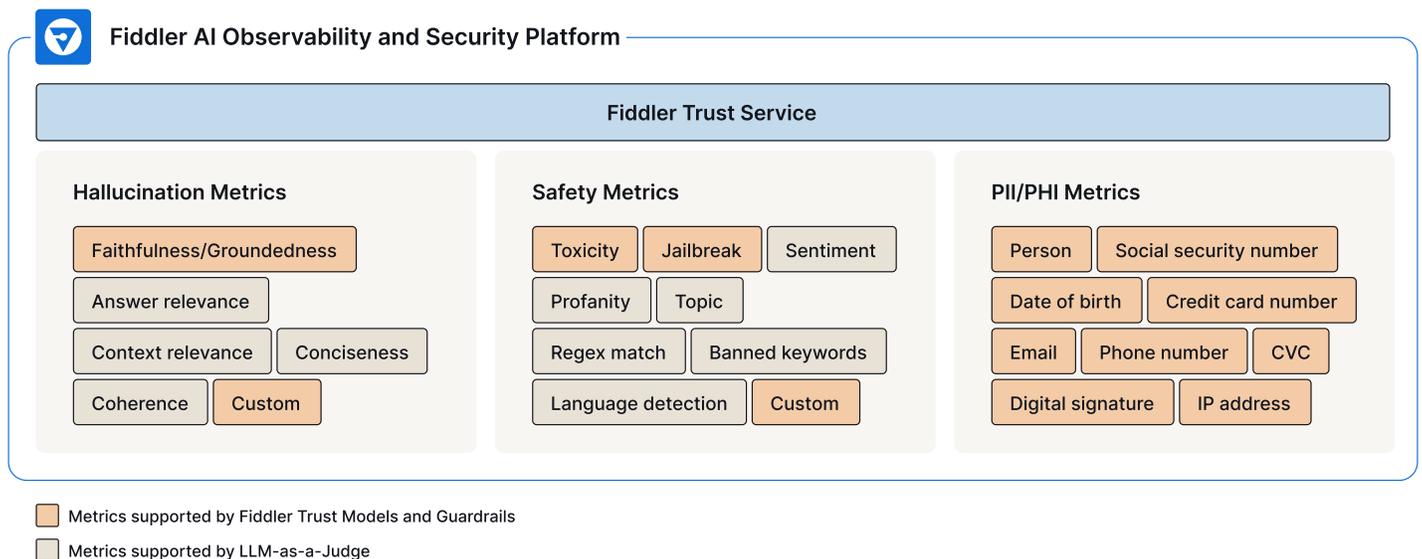
The Fiddler Trust Models are specialized, task-specific models that evaluate agent and LLM prompts and responses at runtime for hallucination, toxicity, jailbreaks, PII/PHI exposure, and other critical risks. Compared to closed source foundational models, they provide low latency, cost effective, and secure scoring for a broad range of LLM metrics.

Enterprises can also create their own metrics by submitting prompts to the Fiddler-hosted Llama 3.1 8B model to address domain-specific use cases and specialized requirements. This fully-managed solution handles 300K+ daily events without the burden of infrastructure management.

Pay \$0 Trust Tax with Fiddler



Fast, Cost-Effective, and Secure Monitoring of 80+ LLM Metrics



Fiddler Guardrails

LLM safety mechanism that moderates risky prompts and responses in real-time before they cause damage.

At <100ms latency, Fiddler Guardrails is the fastest in the industry. It leverages the scoring of the Fiddler Trust Models to evaluate prompts and responses and moderate harmful outputs for hallucination, toxicity, and jailbreaks. Simply specify your desired metric thresholds and let Fiddler Guardrails handle the enforcement.

The Fiddler Trust Service Excels at Popular and Niche Agentic and Generative AI Use Cases

Agentic Systems

Orchestrate workflows with embedded LLM and ML models ensuring accurate, safe task completion.

AI Chatbots

Boost investor value and confidence with accurate financial advice and recommendations from AI chatbots.

Internal Copilot Applications

Enhance employee productivity and boost their confidence in decision-making.

Compliance and Risk Management

Detect adversarial attacks and data leakage.

Content Summarization

Deliver highly accurate summaries for your users.

LLM Cost Management

Increase LLM operational efficiency gains.

Fiddler is the all-in-one AI Observability and Security platform for responsible AI. Our evaluations, monitoring and analytics capabilities provide visibility, context and control across development and production. This gives teams actionable insights to build better, more reliable AI agents, and GenAI and ML applications. An integral part of the platform, the Fiddler Trust Service provides quality and moderation controls for AI agents and GenAI applications. Powered by cost-effective, task-specific, and scalable Fiddler-developed Trust Models — they deliver the fastest guardrails in the industry. Fiddler offers flexibility in secure deployment options through cloud and VPC environments.

Fortune 500 organizations use Fiddler to scale AI agents, GenAI, and ML deployments. This helps them deliver high performance AI, avoid costly AI risks, and maximize ROI.