

ATX Guardrails

The Governance Layer for Enterprise AI Agents

ATX Guardrails provides a unified ecosystem to evaluate agent performance, enforce strict access controls, and filter malicious inputs in real time, effectively shielding your infrastructure while ensuring data privacy. It turns AI governance from policy on paper into controls that sit directly in the request path, so every deployment is defensible to your risk, audit and compliance teams.

Three Controls, One Governance Layer

A Access Control

delivered by **ATX RBAC**

What an agent may touch. Context-aware, role-based filtering at the vector store holds agents to least-privilege scopes and authorized data only, protecting PII and enterprise data privacy.

B Budget Control

delivered by **ATX AI Eval + AI Security Eval**

What an agent may spend. Every run is scored for cost and efficiency, with redundant calls and infinite loops flagged, and denial-of-wallet attacks tested for before they reach production.

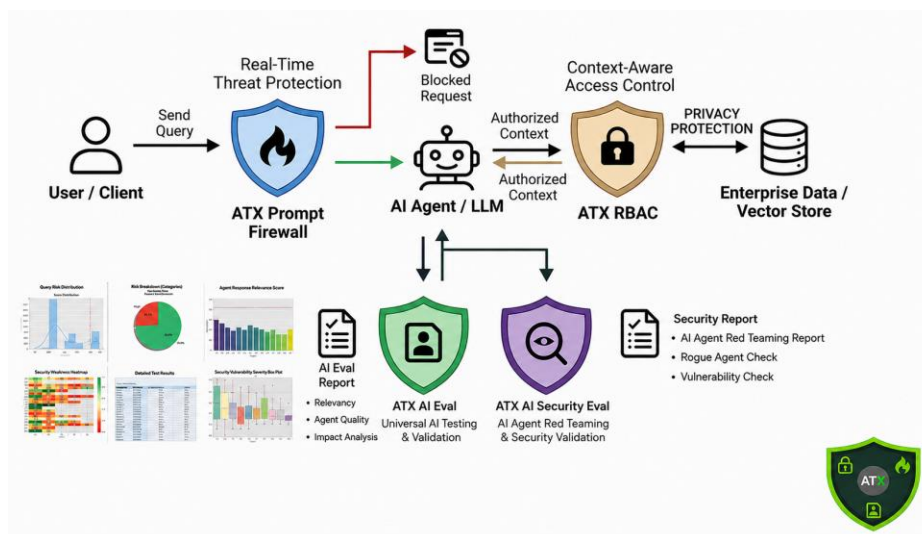
C Control — Human in the Loop

delivered by **observability + approval gates**

Who signs off. Full observability across every agent run: when a check is flagged the process stops rather than proceeds, escalating to a person for review and approval.

What sits inside ATX Guardrails

- **ATX Prompt Firewall** The first line of defense. Inspects prompts to block injections, jailbreaks and policy violations in real time, with pre-configured Safe Rails.
- **ATX RBAC** Context-aware access control. Integrates with vector stores for role-based filtering, so agents retrieve only authorized data.
- **ATX AI Eval** Universal agent testing and validation across quality, tool correctness and production readiness. Also available as a separate service.
- **ATX AI Security Eval **NEW**** Adversarial red-team evaluation for agentic systems. Prompt injection, tool abuse, PII leakage, data exfiltration, denial of wallet and rogue-agent propensity. Mapped to OWASP Top 10 for Agentic Applications and MITRE ATLAS.



How to get started?

```
pip install atx-guardrails
```

Start your enterprise agentic AI journey with governance you can evidence, and confidence you can audit.

