

ATX AI Eval

Evaluate Smarter. Deploy Confidently. Stay Compliant.

ATX AI Eval is a universal evaluation suite for AI agents and LLM applications, available as a standalone offering. It grades agent behavior with a multi-judge LLM jury, attaches to any framework automatically, and generates compliance-ready documentation, so teams move agents into production on evidence rather than guesswork. Delivered as a Python library with a simple API, it is API-key authenticated and logs every evaluation run.

Evaluation metrics

Core quality & correctness	
Response Relevance	Is the answer relevant to the question actually asked?
Task Success	Did the agent complete what it was asked to do?
Tool Correctness	Did the agent call the right tools for the job?
Param Correctness	Were the parameters and arguments passed correctly?
G-Eval	Flexible scoring against any custom criteria you define.
CEP	Conditional branching evaluation. If this, then score that.
Agent-native & production readiness	
CLEAR Score	Composite production-readiness score ($p = 0.83$ with real-world success): Cost, Latency, Efficacy, Assurance, Reliability.
Agent Handoff Quality	How well one agent passes a task to another, task fidelity, context use, goal alignment.
Context Drift	Detects when a multi-turn agent drifts from original user intent, and flags exactly when it started.
Efficiency Score	Whether the agent took the optimal path, flags redundant tool calls and estimates cost waste.
Agent Loop Detector	Identifies agents stuck in infinite cycles before they burn through token budgets.

Why ATX AI Eval?

- **Any framework, automatically.** OpenTelemetry-based instrumentation, no per-framework patching.
- **Three-judge consensus.** Every metric runs through an independent LLM jury, each score surfaced.
- **Conditional evaluation pipelines.** Branching logic defined as configuration, not code.
- **Agent-native by design.** Built for multi-agent systems, tool use and multi-turn sessions.
- **Runs where your data lives.** SaaS, local Docker, in-VPC or fully air-gapped.
- **Audit-ready by default.** Regulation-tagged results, compliance docs and immutable audit trails.

How to get started?

`pip install atx-guardrails`

API-key authenticated, hosted on our server, every evaluation run logged automatically. Start evaluating your agents today and move to production on evidence, not guesswork.