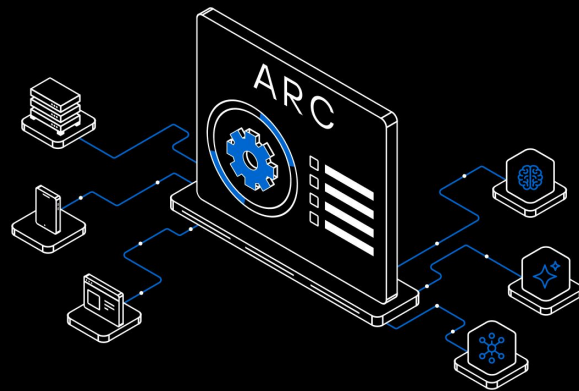


AI Runtime Control (ARC)

Govern your AI traffic and control your spend with an intelligent control plane for the agentic web. Track real-time spend, enforce access controls, and build instant multi-model resilience.



Enforce governance across every agentic interaction

- ✓ Eliminate risks from vibe coding & shadow IT
- ✓ Audit 100% of your AI activity
- ✓ Protect your data and your reputation

Understand AI utilization across your organization

- **Get unified multi-provider metering.** ARC consolidates AI traffic into a single pane of glass with real-time, granular tracking of input and output token counts, request volumes, and calculated costs mapped to virtual keys and specific client sessions.
- **Quickly isolate high-spending hotspots.** Use ARC's real-time dashboards and leaderboards to identify exactly which applications, teams, or developers are consuming the most tokens and dollars.
- **Audit 100% of your AI activity.** Satisfy legal, compliance, and security requirements. Verify how customer data is being handled and trace exactly who requested what.

93%

of organizations
are exceeding AI
budgets

Turn provider outages into a non-event.

- Keep your customer-facing AI features running smoothly through rate limits or upstream provider outages. If a primary model goes down, ARC transparently retries and fails over to your backup models at the edge, ensuring a continuous user experience.

Stop runaway agents from blowing your AI budget or quota.

- Rate limiting per virtual key protects your upstream provider quotas from being exhausted by runaway autonomous agent loops or external traffic spikes.

Learn more at

<https://www.fastly.com/products/ai-runtime-control>