# Magertron™

AI Governance Platform — Request for Proposal Template

Vendor-neutral criteria for governing AI agents, tools, models and the data they reach.

## Purpose

Use this template to evaluate a control plane for agentic AI traffic — not a proxy in front of model providers.

The questions that decide whether you can put agents into production are **who authorized this call, what data did it reach, what did it cost, and can you prove it a year from now.** This template puts those first and model routing where it belongs.

**How to score.** Ask for a live demonstration of policy *enforcement*, not a dashboard. A refusal you can watch — with the reason it names — is worth more than any feature matrix, including this one.

| | |
|---|---|
| ✓ **Supported** | Shipped, in the product today |
| ◗ **Partial** | Some of this; the row says what is missing |
| 🔗 **Integrates** | Delivered through a named third party |
| ○ **Roadmap** | Committed, not shipped |
| — **Not offered** | Deliberate; the row says why |

## 1 Identity and delegation

*The question nobody asks until an auditor does: when an agent acted, who was it acting for?*

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Agent identity** | Does every agent have its own attributable, scoped, revocable identity — never a shared API key? | Must have | ✓ **Supported:** service accounts with per-tool RBAC, revocable without touching a person's account |
| **Delegated calls** | Can an agent act *for* a named person, with both identities carried on the call? | Must have | ✓ **Supported:** RFC 8693 token exchange, plus an on-behalf-of header path for in-cluster callers |
| **Intersection enforcement** | On a delegated call, are BOTH principals evaluated with the narrower winning — or does the agent inherit the person's authority? | Must have | ✓ **Supported:** intersection, never union; delegation can only narrow |
| **Standards compliance** | Is delegation done to a standard your architects can review, or a bespoke header? | Should have | ✓ **Supported:** RFC 8693 against your IdP |

| | | | |
|---|---|---|---|
| **Non-delegable roles** | Can an agent be prevented from acting for an administrator? | Must have | ✓ **Supported:** refused outright with the role named; configurable set |
| **Works without an IdP** | Can delegation work for a deployment that has not finished its SSO rollout, or has none? | Depends | ✓ **Supported:** platform-issued identities are first-class subjects and actors |
| **Mixed-issuer delegation** | Can an IdP-issued agent act for a locally-authenticated person, each verified against its own issuer? | Should have | ✓ **Supported** |

## 2 Authorization and data reach

*Tool permissions are the floor. What the tool can touch is the ceiling.*

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Per-tool authorization** | Is authorization per TOOL, or only per server or endpoint? | Must have | ✓ **Supported:** per tool, per principal |
| **Data-object governance** | Can you govern which tables, schemas and catalogs an agent may reach through a tool — not just which tool it may call? | Must have | ✓ **Supported:** catalog grants over Databricks Unity Catalog, Snowflake, BigQuery |
| **Blast-radius analysis** | Before granting, can you see what a grant would expose? After, what it actually reached? | Should have | ✓ **Supported:** grant simulation and an entitlement-versus-usage graph |
| **Dark entitlements** | Can you see what was granted and never used, and what was used without a grant? | Should have | ✓ **Supported:** both surfaced as distinct states |
| **Tenant isolation** | Are environments isolated such that a scoped administrator cannot read across them? | Must have | ✓ **Supported:** namespace isolation with delegated administration |
| **Revocation latency** | When a grant is revoked, how long until the next call is refused? | Must have | ◗ **Partial:** catalog grants refresh within a configurable window; a ROLE change does not reach an already-issued token |

## 3 Supply-chain integrity

*The tool you approved in March is not necessarily the tool you are calling in September.*

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Definition baselining** | Is each tool's schema recorded at approval and compared on every discovery? | Must have | ✓ **Supported** |
| **Drift enforcement** | When a vendor changes a tool's schema, do calls STOP until a human reviews the diff — or is it logged and allowed? | Must have | ✓ **Supported:** refused regardless of permissions until reviewed |
| **Active detection** | Are servers probed for change, or is drift only noticed on the next call? | Should have | ✓ **Supported:** scheduled probing with on-demand re-probe |

| | | | |
|---|---|---|---|
| **Prompt-injection scanning** | Are tool descriptions scanned for instructions aimed at the model rather than the user? | Should have | ✓ **Supported:** verdict recorded per tool at persist time |
| **Quarantine** | Can a server and all its tools be disabled atomically? | Should have | ✓ **Supported** |

## 4 Cost governance

*"Cost dashboards" show you what you spent. Governance stops you spending it.*

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Attribution** | Is spend attributed to the principal that caused it — including the person an agent acted for — or to the application? | Must have | ✓ **Supported:** both identities on the meter record; delegated calls bill the person |
| **Budgets that refuse** | At the limit, is the call REFUSED — or alerted after the money is spent? | Must have | ✓ **Supported:** refused at the gateway |
| **Contract awareness** | Can prepaid commitments and pay-as-you-go be tracked separately, with a position that does not reset monthly? | Should have | ✓ **Supported:** per-vendor contracts, burn-down reconciled to the vendor console |
| **Multi-dimensional metering** | Beyond tokens — can calls be rated on per-tool rate cards and magnitude dimensions? | Should have | ✓ **Supported** |
| **Chargeback** | Can finance receive a per-department, per-vendor, per-tool statement from the same records security uses? | Should have | ✓ **Supported:** one record answers the auditor and the CFO |
| **Rate limiting** | Can throttles be applied per identity, tool and server? | Must have | ✓ **Supported** |

## 5 Audit and evidence

*Every vendor logs denials. Ask what happens on an allow.*

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Allows recorded** | Are ALLOWS recorded with the rule that permitted them — or only denials? | Must have | ✓ **Supported:** an allow record answers "who read this last quarter, and under what grant" |
| **Authority record** | Can you reconstruct the roles and permissions **in force at the moment of the call**, years later? | Must have | ✓ **Supported:** full authority snapshot per turn, exportable as JSON |
| **Refusal detail** | Does a denial name WHICH check failed, which principal and which rule — or just "denied"? | Must have | ✓ **Supported** |
| **SIEM export** | Do events reach your SIEM in a normalized schema? | Must have | ✓ **Supported:** OCSF-aligned, into Splunk, Elastic or Datadog |
| **OTEL export** | Can traces and metrics export to OTEL-compatible systems? | Must have | ○ **Roadmap:** audit and metering are structured today; an OTEL exporter is committed |

| | | | |
|---|---|---|---|
| **Continuous control evidence** | Is compliance evidence generated continuously, or assembled the week before the audit? | Depends | ✓ **Supported:** continuous SOC 2 control evidence with signed per-turn receipts |
| **Vendor attestations** | Does the VENDOR hold SOC 2 / HIPAA / GDPR attestations? | Depends | ◗ **Partial:** the platform generates control evidence for YOUR audit; the vendor's own attestation is in progress. Ask every vendor which they mean |

## 6 Endpoint reality

*Governance covers what routes through the gateway. This is how you find what does not.*

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Shadow MCP discovery** | Can you see every MCP server configured on managed endpoints — approved or not? | Should have | ✓ **Supported:** reads MCP client configs across Claude Desktop, Claude Code, Cursor, VS Code, Windsurf |
| **Classification** | Are declared servers classified as routed through the gateway, pointed straight at a vendor, or pointed at a route that no longer exists? | Should have | ✓ **Supported** |
| **Remediation** | Can endpoints be re-pointed at the gateway by policy, without touching the laptop? | Should have | ✓ **Supported:** observe / advise / enforce, per namespace |
| **Delivery** | Does it use the endpoint management you already run? | Must have | ✓ **Supported:** Jamf, Intune, Ansible, or a script and a scheduler. No resident agent |

## 7 Model access and routing

*Table stakes. Necessary, not sufficient.*

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Governed model access** | Are model calls subject to the same identity, budget and audit controls as tool calls — or routed around them? | Must have | ✓ **Supported:** the model leg runs through the same gateway, metered and attributed |
| **Provider key isolation** | Can applications call models without ever receiving a raw provider key? | Must have | ✓ **Supported:** credentials injected at the gateway, rotated centrally |
| **OpenAI-compatible endpoint** | Is there an OpenAI-shaped endpoint so existing clients need no change? | Must have | ○ **Roadmap:** committed as a translation layer over the existing governed route |
| **Provider coverage** | Which providers are supported, and how quickly is a new one added? | Must have | ◗ **Partial:** major hosted and self-hosted providers, config-driven onboarding. Not a 1000-model catalog |

| | | | |
|---|---|---|---|
| **Routing and fallback** | Can requests be routed and failed over by model, provider, latency or weight? | Should have | ◗ **Partial:** routing policy and model comparison; latency- and weight-based balancing is roadmap |
| **Model comparison** | Can one prompt be run across models with cost side by side? | Nice to have | ✓ **Supported** |
| **Caching** | Is there prefix or semantic caching to reduce latency and cost? | Nice to have | ◗ **Partial:** prefix caching; semantic caching is roadmap |
| **Batch API** | Is there an asynchronous batch path at batch pricing? | Nice to have | — **Not offered** |

## 8 Guardrails

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Tool-layer guardrails** | Are tool definitions and tool traffic inspected for injection and poisoning? | Must have | ✓ **Supported** |
| **Content guardrails** | Can PII detection and moderation run inline before or after a model call? | Must have | ○ **Roadmap:** the policy hook exists; content inspection is committed as an integration |
| **Third-party guardrails** | Can external providers (Bedrock Guardrails, Azure Content Safety) run in the decision path? | Should have | ○ **Roadmap** |
| **Deterministic enforcement** | When a guardrail fires, is the call refused — or annotated and forwarded? | Must have | ✓ **Supported:** the gateway refuses. Nothing is annotated and let through |

## 9 Deployment and ownership

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Where the control plane runs** | Does the component making authorization decisions run in YOUR cluster, or the vendor's? | Must have | ✓ **Supported:** entirely in your cluster. Nothing calls home |
| **Data residency** | Do prompts, audit records and cost data ever leave your infrastructure? | Must have | ✓ **Supported:** they do not |
| **SaaS option** | Is a hosted option available? | Depends | — **Not offered, deliberately:** the decision path sits between your agents and your tools; putting it across the internet adds a round trip and a dependency to every call |
| **Air-gapped operation** | Can it run with no outbound connectivity at all? | Depends | ✓ **Supported,** including delegated identity without an external IdP |
| **Install** | How long from bare cluster to governed first call? | Should have | ✓ **Supported:** one command; Helm chart for an existing cluster |

| | | | |
|---|---|---|---|
| **Infrastructure as code** | Terraform, Helm, GitOps? | Must have | ◗ **Partial:** Helm and GitOps today; a Terraform provider is roadmap |
| **Multi-cluster** | Can several clusters be governed and reported on together? | Should have | ○ **Roadmap:** federation |
| **Scale** | Demonstrated throughput, and behavior under load? | Must have | ✓ **Supported:** 2,500 rps sustained on real tool calls; Kubernetes-native autoscaling |

## 10 MCP and agent operations

| CRITERIA | WHAT SHOULD YOU EVALUATE? | PRIORITY | MAGERTRON |
|---|---|---|---|
| **Server registry** | Can MCP servers be registered, discovered and governed centrally? | Must have | ✓ **Supported** |
| **Tool discovery** | Are tools auto-discovered and callable from a console or portal? | Must have | ✓ **Supported:** developer portal with a working request against the gateway per tool |
| **Developer self-service** | Can a developer see exactly what THEY may call, and get a working example? | Should have | ✓ **Supported** |
| **Virtual MCP servers** | Can tools from several servers be presented as one? | Nice to have | — **Not offered, deliberately:** composition the caller cannot influence is a naming convenience. Curation is done with per-tool permissions — grant a role one tool from a forty-tool server and that is what the portal, the catalog and the model each see |
| **Agent and skills registry** | Can agents and reusable skills be published, discovered and governed? | Should have | ○ **Roadmap** |
| **CLI and API** | Is everything the console does available on an API and a CLI? | Must have | ✓ **Supported** |

# Questions to ask every vendor, including us

Independent of the matrix. The answers are more revealing than the checkmarks.

1. **Show me a refused call.** Not a dashboard — a refusal, and the reason it gives. Does it name which check failed, which principal and which rule?
2. **Show me an allow record.** Then ask it to reconstruct the permissions that were in force when that call happened.
3. **An agent acts for a person. Show me both identities on the record** — and what happens when the person's permissions are narrower than the agent's.
4. **A tool's vendor changes its schema overnight.** What happens to the next call?

5. **Where does the component that authorizes calls run,** and what data leaves my perimeter?
6. **What on this sheet would you mark partial?** A vendor who cannot name one is either not listening or not telling you.