ARIA
Responsible AI, Verified.

Aggi L.L.C.

AI Endpoint Onboarding Guide

Getting your LLM connected, tested, and continuously monitored

Prepared by Aggi Technologies LLC · aria.aggicorp.com · prasad@aggicorp.com

This guide walks you through connecting your AI system to ARIA and activating the three layers of continuous oversight ARIA provides: **behavioral testing** (does your AI follow its own policy?), **runtime guardrails** (blocking policy violations in real time), and **drift monitoring** (alerting you when your AI's behavior changes). By the end you will have a working connection, your first evaluation results on screen, and a live alert threshold set.

If you decide ARIA is a good fit after working through this, a subscription converts your pilot directly into ongoing continuous compliance — no re-configuration needed.

Before you start

You need:

- Your ARIA login URL and credentials (provided by Aggi Technologies on engagement start).
- An API key for the LLM you want to connect (Anthropic, OpenAI, AWS, Google, Azure, or self-hosted).
- Your role in ARIA must be **Org Admin** or higher to register endpoints.

> **On cost.** ARIA does not absorb your LLM provider's token charges. When ARIA runs a test against your endpoint, those tokens appear on your own provider bill — not ours. A standard evaluation run costs $0.01–$0.05 on Sonnet-class models. ARIA's platform fee covers orchestration and analysis; the inference bill stays yours.

Step 1 — Complete the onboarding wizard (10 minutes)

If your organization is new to ARIA, a banner on the Dashboard invites you to start the onboarding wizard. The wizard takes five questions and results in ARIA recommending the right compliance framework set for your industry and regulatory situation.

The five questions ask about:

- Your industry
- How your AI system is deployed
- Which geographies it operates in

- What kind of data it touches
- Your current audit posture

ARIA's rule engine — not an LLM — maps your answers to a framework recommendation deterministically. You can override the recommendation before committing.

Prospect type	Typical framework set
US healthcare / medical AI	NIST AI RMF · HIPAA · FDA CDS · ISO 42001
EU medical device / SaMD	EU MDR · EU AI Act · ISO 42001
Hospital / telehealth	HIPAA · SOC 2 · HITRUST
Pharma / drug discovery	NIST AI RMF · ISO 42001 · EU AI Act (no HIPAA)
US federal contractor	FedRAMP Moderate · NIST AI RMF
EU financial services	DORA · SOC 2 · GDPR
Enterprise SaaS (B2B)	ISO 42001 · SOC 2

Step 2 — Register your LLM endpoint (5 minutes)

Navigate to **Admin** → **LLM Endpoints**. Click **Register endpoint**.

What to fill in

Name.

Something human-readable that identifies the system — for example *VisitTriage Symptom Router (prod)* or *ChartScribe Ambient (staging)*. This name appears in evaluation results and drift alerts, so make it specific.

Provider.

Select the cloud that hosts your model.

Endpoint URL or region.

Provider	What to enter
Anthropic	https://api.anthropic.com
OpenAI	https://api.openai.com
Amazon Bedrock	Your AWS region, e.g. us-east-1 (not a URL)
Google Vertex AI	Your GCP region, e.g. us-central1 (not a URL)
Azure OpenAI	https://.openai.azure.com

Self-hosted	Base HTTPS URL of your endpoint, e.g. https://vllm.example.com/v1 — do NOT include /chat/completions

API key.

Paste your provider key. ARIA encrypts it immediately on save — it is never readable again after registration, including by Aggi staff.

Provider-specific extras.

Bedrock requires your AWS secret-access key in addition to the access-key ID. Vertex requires a GCP service-account JSON blob. Azure requires your deployment name. ARIA's form prompts you for these.

Capabilities.

Check Chat, Embedding, and/or Vision as applicable. Most evaluations use Chat.

Self-hosted endpoints

If you run your own LLM (vLLM, llama.cpp, Ollama, Text Generation Inference, or any server that speaks the OpenAI chat-completions protocol), select **self_hosted**. You must also pin the model name in each evaluation fixture, because ARIA cannot guess which model is running.

> ARIA refuses plain HTTP URLs for self-hosted endpoints. Your endpoint must be TLS-terminated. If it is not yet, put an nginx or Caddy reverse proxy in front before registering.

Step 3 — Run your first evaluation (5 minutes)

With your endpoint registered, find its row on **Admin** → **LLM Endpoints** and click **Run eval**. ARIA asks you to confirm — it will tell you how many fixtures will run and the estimated token cost. Confirm. The run completes in roughly 30 seconds.

Navigate to ***/eval*** to see the results under “Your organization’s eval posture.”

What you are looking at

Each evaluation run fires a set of behavioral test fixtures against your LLM. A fixture is a structured prompt that should produce a specific kind of response — a clinical disclaimer, a refusal to echo back PHI, an acknowledgement that the AI is not a doctor. ARIA scores each response pass or fail.

The fixture set ARIA runs depends on which frameworks you selected. The total catalog covers **160 behavioral fixtures across 12 frameworks**.

Framework	Fixtures	Sample invariants tested
NIST AI RMF	22	Harmful-output refusal, explainability, fairness, prompt-injection, hallucination

FDA CDS	30	Clinical hallucination, PHI refusal, drug-interaction caution, contraindication check
ISO 42001	25	Human oversight path, training-data transparency, audit-logging disclosure
EU AI Act	15	Art. 5 prohibitions, Art. 9 risk management, Art. 13 transparency, Art. 14 human oversight
SOC 2	10	PII handling discipline, least-privilege enforcement, output accuracy caution
GDPR	10	Art. 22 human-review, right-to-erasure acknowledgement, data minimization
HITRUST	10	PHI redaction, cross-session PHI refusal, audit log integrity
CCPA/CPRA	10	Automated-decision disclosure, right-to-delete acknowledgement
EU MDR	10	CE marking obligations, ISO 14971 risk, PMCF monitoring, UDI disclosure
DORA	10	ICT risk governance, incident classification, third-party AI resilience
FedRAMP	8	Change control, incident plan, input validation, access control
HIPAA	0	No Shape B fixtures — covered by assessment questionnaire

What a fail means

A failing fixture tells you that a specific policy invariant was violated on this run. It does not mean your model is broken — it means there is a gap between what your policy says and what your model did when tested. That gap is exactly what ARIA is designed to find.

> Click any run row to drill into the per-fixture grid. For each fixture you will see the prompt, the model's response, and whether it passed or failed. Compliance Officers and above see the underlying LLM call record; other roles see the outcome without the raw text.

Step 4 — Schedule ongoing evaluations and set a drift threshold (5 minutes)

A single run is a data point. What ARIA is built for is the trend. On the endpoint row, click **Schedule**.

Setting	Guidance
Cadence	Weekly is the standard starting point. Daily for production endpoints in regulated environments. Monthly for staging systems.
Hour (UTC)	Pick an off-peak hour. ARIA runs the fixtures and your provider charges the tokens.
Drift alert threshold	A percentage. If the rolling pass-rate moves by more than this versus the 30-run baseline, you get an alert. 10% is a reasonable starting point.

Click **Save schedule**. ARIA now runs your evaluation on the cadence you set and compares each run against the rolling baseline.

Step 5 — Activate the inline proxy (optional, 15 minutes)

The evaluations in Step 3 test your LLM on a schedule. The inline proxy tests every call in real time and can block policy violations before they reach your users.

> **This step is optional during your pilot.** Activate it when you want runtime enforcement, not just periodic testing.

How it works

Instead of your application calling your LLM provider directly, it calls an ARIA URL. ARIA receives the request, runs it through a set of content gates, forwards it to your provider if it passes, runs the response through a second set of gates, and returns the result. The entire round-trip adds less than 15 milliseconds of overhead on pattern gate checks.

Your application does not change its behavior — the ARIA proxy speaks the same OpenAI-compatible protocol your application already uses. **You change one URL and one API key.**

> **Supported providers for the inline proxy:** OpenAI and self-hosted (any server speaking the OpenAI chat-completions protocol). Anthropic, Bedrock, Vertex, and Azure OpenAI are supported for evaluation (Shape B) but not yet for the inline proxy — contact Aggi Technologies for current availability.

Turning it on

- On **Admin** → **LLM Endpoints**, click **Edit** on your endpoint.
- Scroll to the **Inline proxy (Shape C)** section.
- Check **Enable inline-proxy URL for this endpoint** and save.
- Copy the per-endpoint proxy URL that now appears.

Getting your proxy token

- Navigate to **Admin** → **API Tokens**.
- Click **New token**, check the **proxy:invoke** scope, and save.
- Copy the token — it is displayed once only.

Updating your application

Point your existing LLM SDK at the ARIA proxy URL and pass your ARIA proxy token as the API key. That is the complete change. For example, using the OpenAI Python SDK:

```
import openai

client = openai.OpenAI(
    api_key="<your-aria-proxy-token>",
    base_url="https://aria.aggicorp.com/proxy/v1/<your-endpoint-id>/"
)

response = client.chat.completions.create(
    model="gpt-4o-mini",
    messages=[{"role": "user", "content": "Summarize the HIPAA Privacy Rule."}]
)
```

If a gate fires and blocks a call, the response comes back as HTTP 200 with finish_reason: content_filter. Check choices[0].finish_reason to detect a block. The X-Aria-Verdict response header is always block on a blocked call and pass otherwise.

What the gates check

ARIA runs ten gates on every proxied call — seven deterministic pattern gates and three AI-judge gates.

Pattern gates (under 5 ms, fail-open by default)

Pre-call gates run before ARIA sends the request to your provider.

Gate	What it catches
Prompt injection	Attempts to override system instructions or escape role via prompt manipulation
PII classifier	Phone numbers, email addresses, SSNs, credit card numbers, IP addresses in user messages
Jailbreak heuristic	DAN-style activation attempts, developer-mode prompts, fictional-story reframing
Topic restriction	Queries outside the topic boundary you define for the endpoint

Post-call gates run on your provider's response before it is returned to your application.

Gate	What it catches
PII leak	PII categories appearing in the model's response
Hallucination marker	Medical, legal, or financial claims without appropriate hedging language
Output format	Responses that do not match the structured format you require

AI-judge gates (one additional LLM round-trip each)

Judge	What it detects	Default

Hallucination judge	Claims not supported by the context you provided	On
Toxicity judge	Harmful, abusive, or hate content (blocks above threshold)	On
Factuality judge	Claims contradicting a ground-truth snippet you configure	Off — requires configuration

Judges use your registered LLM provider, billed to your account. Judge failures are fail-open — if a judge call errors or times out, the response passes through unchanged.

Gate policy options

Option	What it does
Rate limiting	Cap proxy calls per minute. Calls over the limit receive HTTP 429. Protects against runaway loops or token abuse.
Replay detection	Reject identical requests within a 5-minute window. Useful where double-send or replay attack is a concern.
Fail-closed mode	Block on gate errors instead of passing through. Run at least one week fail-open before enabling — a misconfigured gate will block all traffic.
Custom patterns	Add org-specific regex to prompt-injection, PII classifier, or jailbreak-heuristic gates. For example: internal record identifier formats ARIA's built-in classifier doesn't recognize.

Viewing your results

The Eval page (/eval)

Your organization's evaluation posture lives here. Per-endpoint pass-rate trends over the rolling 30-day window, run history, and the per-fixture grid for any individual run. If a fixture fails consistently across multiple runs, that is a reliable signal — not noise.

ARIA also shows its own evaluation posture on the same page under “ARIA’s vendor posture.” This is ARIA dogfooding its own platform, run nightly. It serves as a reference for what a healthy evaluation surface looks like.

> The same data is publicly available without login at **aria.aggicorp.com/eval-public** — share that URL with your QA team or an auditor who needs to verify ARIA's AI hygiene without creating an account.

The Drift Alerts page (/eval/drift-alerts)

Every time a scheduled re-run moves the pass-rate by more than the threshold you set, ARIA opens a drift alert here. Each alert shows the endpoint, the drift percentage, the observed rate versus the baseline, and when the alert fired.

Click **Acknowledge** to record that someone is investigating, and add investigation notes. Click **Resolve** when the underlying behavior has been corrected. ARIA tracks these transitions in the immutable audit log so you can show an auditor exactly when you detected a problem and when you fixed it.

The Continuous-Compliance Cockpit (/eval/cockpit)

The single-screen overview for compliance officers. Shows all active endpoints, current schedules, baseline-versus-observed pass-rates as sparklines, trend direction, and open alert counts. Refreshes every 30 seconds.

The Dashboard

The higher-level picture: the audit-readiness composite score (0–100, with letter grade), the posture trajectory chart (90-day per-framework trend), the Pulse heat map (180 days of compliance activity), and the Action Queue ranking the open items that would move your score the most.

Band	Score	Meaning
At-risk	0–39	Material gaps. Immediate action required.
Needs Attention	40–69	Partial posture. Work to do before next audit.
Ready	70–100	Defensible posture. Maintain and monitor.

What a weekly evaluation cycle looks like

When	What happens
Monday morning	ARIA sends a weekly compliance summary email. Review in two minutes — if there are no open alerts, everything is tracking normally.
When a drift alert fires	You receive a notification. Open /eval/drift-alerts. Acknowledge the alert and investigate: did your provider update the model? Did your prompts change? Did query patterns shift?
Before an audit	Open /eval and export the run history. Open the Dashboard and export the posture trajectory. These two artifacts answer most of what an auditor will ask.

Common questions during setup

What if my endpoint is not public-facing?

ARIA's orchestrator calls your endpoint from its production infrastructure. Your endpoint needs to be reachable over HTTPS. If it lives behind a private network, whitelist ARIA's egress IP range (provided by Aggi Technologies on request) or use a self-hosted endpoint that bridges the network boundary.

Can I test a staging endpoint without affecting production results?

Yes. Register a separate endpoint for staging, give it a different name, and run evaluations against it. Staging and production results are kept separate in ARIA — they do not pollute each other's drift baselines.

What if my model fails a lot of fixtures on the first run?

That is normal, and it is the point. ARIA's first run is your baseline — it tells you where you actually are, not where you think you are. The pass rates improve over time as you address the gaps the fixtures surface. The drift alert threshold is relative to your baseline, so a low baseline does not trigger alerts until you improve and then slip back.

Can we write our own fixtures?

Yes. Upload a custom fixture YAML under Admin → Custom Fixtures. If your fixture's ID matches a bundled ARIA fixture, your version overrides the bundled one for your organization only. Fixture format documentation is available from ARIA support.

Do the AI-judge calls send our data to Anthropic?

No. The judge calls go to your own registered LLM provider using your own API key. ARIA does not route them to an Anthropic endpoint on your behalf. The data stays within your existing provider relationship, billed to your own account.

How do we share proof with our auditor?

Generate a Posture Summary report from Assessments → [your assessment] → Generate report. This produces a PDF covering your framework posture, open findings, evidence completeness, and the trajectory trend. For ongoing auditor access, share the /eval-public URL — it requires no login and shows ARIA's own evaluation discipline as a trust artifact.

Next steps

At the end of your pilot you will have:

- One or more registered LLM endpoints
- A baseline evaluation result with per-fixture pass/fail data
- A scheduled re-run cadence with a live drift threshold
- Optionally, an inline proxy enforcing gates on every production call
- A Posture Summary PDF you can share with your board or an auditor

If ARIA has been useful, the same configuration carries forward into an ongoing subscription — nothing needs to be rebuilt. Aggi Technologies will work with you on a pricing proposal that fits your endpoint count, evaluation frequency, and framework requirements.

To proceed, or to ask questions during the pilot:

Dr. Prasad N. Golla — Founder, Aggi Technologies LLC
prasad@aggicorp.com · (972) 351-3342 · aria.aggicorp.com