Minsky™ Gateway
MULTI-MODEL AI ROUTING

AI LABS, INC.
SOVEREIGN ENTERPRISE AI

Ai Labs

MINSKY™ GATEWAY

Sovereign AI Routing Platform

A Minsky™ Platform Suite Product | Ai Labs, Inc.

ON-PREMISES · ZERO DATA EGRESS	PDPL · SDAIA · SAMA ALIGNED
OPENAI & ANTHROPIC COMPATIBLE	FULL AUDIT TRAIL

OVERVIEW

Minsky™Gateway is an enterprise AI routing platform that sits between your applications and every major LLM provider. It classifies each request, **convenes a panel of models to deliberate on the best-fit provider**, and dispatches to the winner — independently for every single API call, deterministically, and with a full audit trail. Minsky™Gateway is built and already deployed on AWS infrastructure today.

Built for teams that need governance, not just speed, Minsky™Gateway delivers OpenAI- and Anthropic-compatible ingress, versioned routing policies, per-call model deliberation, HITL approval workflows, real-time cost tracking, board-ready run reports, and compliance tooling aligned to **PDPL, SDAIA, and SAMA** — all deployable on-premises for zero data egress.

PLATFORM MODULES

Dashboard	Sandbox	Model Catalog	Routing Policies
Monitoring	Benchmarks	Cost & Usage	API Key Mgmt
Approvals (HITL)	Compliance & SBOM	Audit Log	

KEY CAPABILITIES

01 Panel-Based Voting on Every Call

Minsky™Gateway doesn't route on price or popularity. For each individual API call, a fast orchestrator classifies the request and nominates a best-fit model; every eligible model in your catalog then casts its own reasoned vote on that choice; and two independent referees — one from OpenAI, one from Anthropic — read every ballot and issue the final ruling. Cross-provider referees remove single-vendor bias, and the entire deliberation runs fresh on each request — nothing is cached, batched, or assumed. It is the same jury-style decision process on every one of thousands of daily calls, giving you a consistent, explainable answer to ‘why this model, for this request.’

02 Intelligent Multi-Provider Routing

Classifies every request by task type, then scores all eligible models on cost, latency, and reliability using configurable policy weights — selecting the best fit for the job, not just the cheapest or most powerful model available.

03 OpenAI & Anthropic Compatible Ingress

Drop-in replacement endpoints for /v1/chat/completions and /v1/messages let existing SDK clients route through Minsky™Gateway with zero code changes.

04 Per-Key Routing Controls

Every API key routes Auto (full classify → panel-vote → referee pipeline) or Manual (pinned to an approved allow-list, cheapest-that-fits, no LLM overhead). No-limit keys available for trusted workloads.

05 Versioned Routing Policies

Define capability filters, cost-tier ceilings, latency budgets, and scoring weights per account; every policy change is versioned and auditable, so you always know exactly which policy was live for any given call.

06 Real-Time Cost & Usage Tracking

Per-request token metering, credit balance enforcement, and time-series cost breakdowns by model and provider keep AI spend visible and under control.

07 HITL Approvals & Compliance Artifacts

Built-in human-in-the-loop approval workflows, a live compliance scorecard, CycloneDX SBOM export, and NDJSON SIEM streaming support enterprise governance requirements out of the box.

08 Board-Ready Run Reports

Export any run — every call a codegen job or agent makes — as a single consolidated PDF: executive summary, model-mix and cost charts, a call-by-call snapshot, the full voting trace for each decision, and a comparison against using any single model for everything.

HIGHLIGHTS

Each API call gets its own jury: a full panel of models votes with written reasoning, and two cross-provider referees rule on the outcome. Every ballot and rationale is logged — so you can see exactly why any model was chosen, on any call.

Deterministic, Explainable Routing

Every routing decision is scored, ranked, and logged with its candidate table, classifier output, panel ballots, and referee reasoning — giving you a reproducible audit trail for every AI call.

Sovereign by Design

Deploy fully on-premises with zero data egress: per-provider data-residency controls, Fernet-encrypted credentials, and hashed API keys ensure your data never leaves your infrastructure.

One Gateway, Every Provider

Register OpenAI, Anthropic, DeepSeek, xAI, Perplexity, and self-hosted models in a single catalog — then let smart policies and fallback chains handle provider selection transparently.

Spend Less, Prove It

Because Minsky™Gateway sends routine work to lighter models and reserves premium models for calls that need them, every run report shows exactly how the workload was allocated.

WHO IT'S FOR

Platform & ML Engineering Teams

Replace fragile per-provider integrations with a single, policy-driven gateway that handles routing, retries, fallback chains, and streaming automatically.

FinOps & AI Cost Owners

Get per-request token metering, credit caps, and model-level cost breakdowns to forecast and control AI spend across every team and application.

Compliance & Risk Officers

Satisfy PDPL, SDAIA, and SAMA requirements with an immutable audit trail, per-call decision records, HITL approval gates, and exportable compliance reports.

Enterprise Developers

Self-register a tenant, grab an API key, and call any provider through a familiar OpenAI- or Anthropic-compatible interface — no provider onboarding required.

MINSKY™GATEWAY vs. OPENROUTER — SNAPSHOT

Dimension	Minsky™Gateway	OpenRouter
Primary audience	Enterprise & government	Individual developers & startups
Routing logic	Deterministic, panel-deliberated, and logged	Task classification + usage popularity
Explainability	Full ballot + referee reasoning logged per call	Not reproducible — “black box” ranking

Bias control	Two cross-provider referees rule on every decision	Single-vendor ranking logic
Auditability	Immutable, versioned, SIEM-ready audit trail	Minimal or none
Deployment	Native on-premises / air-gapped, zero data egress	Cloud-only
Governance / HITL	Enterprise-grade HITL approval workflow	Not built for enterprise governance
Ingress compatibility	Drop-in OpenAI- & Anthropic-compatible endpoints	OpenAI-compatible endpoint only

WHY CHOOSE MINSKY™GATEWAY: SECURITY & TRUST

- **Encrypted credentials, hashed keys** — Provider credentials encrypted at rest with Fernet symmetric encryption; API keys stored as keyed hashes with encrypted recovery for authorized re-issuance — plaintext never exposed in transit logs.
- **RBAC-enforced admin console** — Role-checked on every mutation with IP-stamped audit entries capturing before-and-after state.
- **Immutable audit trail** — For all admin actions and routing decisions, including the full voting record for every call; filterable by action type, target, and actor.
- **HITL approval workflow** — Gates high-risk policy changes with reviewer sign-off before execution.
- **NDJSON SIEM export** — Audit and routing logs stream directly into Splunk, Elastic, or Microsoft Sentinel.

Jury	2	5-Tier	Live
A model jury — full panel + referees on every call	Cross-provider referees rule every routing decision	Low / medium / high / xhigh / max cost-tier ceilings	Built & deployed on AWS today

SEE MINSKY™GATEWAY IN ACTION

Request a live demo or deploy the on-premises trial today — contact your Ai Labs account team.

Ai Labs, Inc. | 430 W Warner Rd, Suite 113, Tempe, AZ, USA | www.ailabsinc.ai

** Minsky™Gateway has been developed and deployed on AWS. Specifications are subject to change.*