AI · TEMPLATE

Do You Need an LLM *Gateway*? Self-Hosted vs. Managed for Regulated AI

KAANSYSTEMS.COM/LIBRARY/LLM-GATEWAY-REGULATED-AI · SEPTEMBER 16, 2026

ABOUT THIS TEMPLATE

An LLM gateway is where you enforce your AI compliance boundary. For PHI, self-host it in-VPC or use Bedrock with a BAA; managed is fine for lower stakes.

THE TEMPLATE

Two parts: decide whether and where the gateway lives, then confirm the controls actually run there. Score each item honestly. Any `no` in the control section blocks regulated traffic through that path until it is resolved.

Do you need one? (gateway trigger)

- [ ] More than one or two features send data to a model provider
- [ ] Any feature sends PHI, PII, or other regulated data into a prompt
- [ ] Provider keys currently live in more than one service or secret store
- [ ] You cannot today answer "what model saw what, when" from one place
- [ ] Two or more of the above true → stand up a gateway; regulated data true → it must be in-boundary

Where does it live? (build-vs-buy)

- [ ] Regulated data in prompts → self-host in-VPC, or use Bedrock under your BAA
- [ ] De-identified / low-stakes only → managed gateway acceptable if disclosed + scored
- [ ] Already on AWS and want managed → evaluate Bedrock before a third-party SaaS gateway
- [ ] If self-hosting: a named owner can deploy, patch, scale, and get paged on the proxy
- [ ] If managed: vendor scored on the vendor-risk rubric, on the sub-processor list, DPA signed (BAA if PHI)

Controls that must run at the gateway

- [ ] PII/PHI redaction runs *before* egress, inside your boundary (not on a vendor's side for PHI)
- [ ] Egress allowlist enforces BAA/DPA-covered providers only — unapproved destinations refused
- [ ] Provider keys held once, in-boundary; rotated centrally; downstream services hold internal keys only
- [ ] Every call emits a content-free trace (ids, versions, hashes, flags, decision) to your own evidence store
- [ ] Prompts/responses are NOT logged as content to the gateway's or a vendor's default dashboard
- [ ] Model versions pinned (dated snapshots, not moving aliases) and recorded per request
- [ ] No silent fallback to a hosted tier that would route prompts back outside the perimeter

Prove it

- [ ] The one place data is scrubbed, keys rotate, and calls are logged is named and documented
- [ ] You can state, per feature, whether that place is inside your boundary or a vendor's
- [ ] The gateway (self-hosted or managed) is in scope for the next access review + evidence pipeline
- [ ] A re-review triggers when a provider is added, a vendor changes terms, or the data flow changes

Most SMBs find the trigger section mostly true by their third or fourth AI feature and land on a self-hosted proxy in-VPC or Bedrock for anything touching regulated data, with a managed gateway reserved for the de-identified lane. All three shapes are defensible for the right data class. The one indefensible position is having no chokepoint at all once the AI features multiply — because that is the position where you cannot answer where the data went.

WHAT IS AN LLM GATEWAY?

An LLM gateway is a reverse proxy in front of your model providers that centralizes the cross-cutting concerns every AI feature needs: authentication and per-app internal keys, routing and failover across providers, rate limiting, cost and token tracking, request/response logging, PII/PHI redaction, and a single provider-agnostic API. Instead of each service holding its own OpenAI or Anthropic key and implementing its own logging, every service calls the gateway, and the gateway calls the provider.

It is the API-gateway pattern applied to model calls. The value is consolidation: one place that knows how to talk to every provider, one place where a prompt can be inspected before it egresses, one place where a model call is recorded, and one place where a leaked key can be rotated without a redeploy of six services. In a system with a single AI feature you may not feel the need. By the fifth feature — a summarizer, a classifier, a draft-generator, a support-bot, an embedding pipeline — the alternative to a gateway is the same four controls implemented five times, slightly differently, with one of them wrong.

Popular self-hostable options include LiteLLM and Portkey's open-source proxy; managed and SaaS gateways, plus the model platforms' own gateway features, cover the same surface as a service. Feature sets and pricing

move fast in this category, so treat any specific capability as list-and-approximate and verify the current split before you commit.