SCORED.TOOLS · ROUNDUP

Best AI API Gateway Tools for LLM Apps in 2026

We tested the top AI API gateways for LLM applications in 2026. Here's how Portkey, LiteLLM, OpenRouter, Kong, Helicone, and others actually stack up.

Why you need an AI gateway in 2026

If you're shipping anything with LLMs in it, you already know the pain: one provider goes down, costs balloon overnight, prompts drift across environments, and nobody on the team can tell you which model handled which request. *AI gateways are the infrastructure layer that sits between your application and the model providers. They handle routing, cost control, authentication, caching, and observability for AI traffic.*

The market has matured fast. There are now serious differences between gateways built for AI from day one and traditional API management tools that bolted on LLM support later. This roundup ranks the eight we've actually used in production, in order of how well they hold up under real workloads.

1. Portkey — Score: 9.4/10

Portkey is the most complete purpose-built AI gateway on the market right now. *It's a purpose-built AI infrastructure platform focused on getting LLM applications to production, providing a unified interface to 250+ models with deep observability, prompt management, and guardrails.* The MCP support matters more than most teams realize: *the MCP Gateway became generally available in January 2026 with MCP protocol support for agent workflows.* And if you have compliance requirements, *Portkey's gateway went fully open source in March 2026, and Enterprise customers can self-host with hybrid or air-gapped deployment options.*

Best for: Production LLM apps and agent workflows where observability, guardrails, and multi-provider routing all matter.

Pricing: Free open-source core; managed tier starts free with paid plans for teams; enterprise pricing on request.

Caveats

- *Some users report the platform can be overwhelming for new users due to the vast array of features.*
- *Enterprise pricing is complex — key features like budget limits are restricted to Enterprise customers only.*

2. LiteLLM — Score: 9.1/10

Litellm is the default choice if you want to self-host and stay in control. *It acts as a proxy that mimics the OpenAI API format while letting you connect to many providers or even other gateways. You can run it inside your own environment, giving you control over data, cost, and deployment. This makes it especially useful for teams working with private models or strict compliance requirements. LiteLLM supports 100+ providers behind a single OpenAI-compatible API, handles fallbacks and budget controls, and runs free on any VPS.*

Best for: Engineering teams who want an open-source, self-hosted gateway they can bend to their infrastructure.

Pricing: Free (open source); LiteLLM Cloud and enterprise support tiers available.

3. OpenRouter — Score: 8.7/10

Openrouter is almost always the fastest way to get started. *OpenRouter is the fastest way to access multiple LLM providers. If you want to prototype with different models or run a small-to-medium workload without managing infrastructure, it's the obvious choice. At scale, the 5.5% fee starts to matter*. It shines when you want unified billing and instant access to hundreds of models without wiring up API keys for each provider.

Best for: Prototyping, side projects, and teams that value simplicity over granular control.

Pricing: Pass-through provider pricing plus a ~5.5% platform fee; no monthly subscription.

4. Cloudflare AI Gateway — Score: 8.3/10

Cloudflare Ai Gateway is compelling if you're already on Cloudflare's edge. It plugs cleanly into Workers, gives you caching and rate limiting at the edge, and the analytics dashboard is genuinely useful. It's less feature-rich than Portkey on prompt management and guardrails, but for teams optimizing for latency and cost-per-request on a global footprint, it's hard to beat.

Best for: Teams already running on Cloudflare who want edge caching, rate limiting, and basic observability.

Pricing: Free tier with generous limits; paid tiers bundled with Cloudflare Workers.

5. Helicone — Score: 8.1/10

Helicone focuses hard on observability rather than trying to be everything. *It provides a unified OpenAI-compatible API for 100+ LLM providers, combines gateway routing with request-level observability for latency, token usage, and cost, supports automatic fallbacks across providers when an API fails or rate limits traffic, offers provider routing including cheapest-available routing and switching when providers hit rate limits or outages, and includes cost tracking and optimization tooling across providers and models.* If your main pain is "we can't see what's happening in production," this is the shortest path to a fix.

Best for: Teams whose primary need is deep request-level observability and cost analytics.

Pricing: Free tier available; paid plans scale with request volume.

6. TrueFoundry AI Gateway — Score: 7.9/10

Truefoundry targets enterprise AI teams and shows it. *It offers enterprise reliability with SOC2, ISO27001, HIPAA, and GDPR compliance, with SaaS, hybrid, and air-gapped deployment options, and a 99.99% uptime SLA*. The tradeoff is that it's overkill if you're a small team just trying to route between OpenAI and Anthropic. Pricing and onboarding both lean enterprise.

Best for: Regulated industries and enterprise teams that need air-gapped or hybrid deployments.

Pricing: Enterprise pricing on request; no self-serve tier.

7. Vercel AI Gateway — Score: 7.6/10

Vercel Ai Gateway is the natural fit if you're already building on Vercel with the AI SDK. It handles provider fallbacks, unified billing, and observability with almost zero configuration if your app is a Next.js deployment. It doesn't try to be a full LLMOps platform — no prompt management or advanced guardrails — but for the target audience, that's fine.

Best for: Next.js and Vercel-hosted apps that want a drop-in gateway with minimal setup.

Pricing: Usage-based, bundled into Vercel plans.

8. Kong AI Gateway — Score: 6.9/10

Kong Ai Gateway is the honest "only if" pick. *Kong AI Gateway makes sense if and only if your organization already runs Kong. Adding LLM routing to your existing API management layer is smarter than deploying a separate gateway. But don't adopt Kong just for LLM routing, that's like buying a tractor to mow your lawn. Kong's plugin ecosystem was built for REST APIs, not LLM*

workloads. The AI-specific additions — semantic routing, token-based rate limiting, load balancing — are layered onto a platform not designed with LLM-native requirements in mind.

Best for: Organizations already standardized on Kong for API management.

Pricing: Open-source core; enterprise pricing on request.

Comparison table

Tool	Score	Self-host	Model count	Best for
Portkey	9.4	Yes (OSS + Enterprise)	250+	Production LLM + agent apps
LiteLLM	9.1	Yes (OSS)	100+	Self-hosted control
OpenRouter	8.7	No	300+	Prototyping, unified billing
Cloudflare AI Gateway	8.3	No	Major providers	Edge + Cloudflare shops
Helicone	8.1	Yes	100+	Observability-first teams
TrueFoundry	7.9	Yes (hybrid/air-gapped)	100+	Enterprise + regulated
Vercel AI Gateway	7.6	No	Major providers	Next.js / Vercel apps
Kong AI Gateway	6.9	Yes	Major providers	Existing Kong users only

Final picks

- **Best overall:** Portkey — the most complete AI-native gateway with MCP, guardrails, and observability in one place.
- **Best open-source / self-hosted:** Litellm — battle-tested, easy to deploy, and you own the whole stack.
- **Best for prototyping:** Openrouter — five minutes to first request, no infrastructure to manage.
- **Best for observability:** Helicone — if you just need to see what's happening, start here.

- **Skip unless you already use it:** Kong Ai Gateway — great API gateway, awkward AI gateway.

Pick the one that matches where you actually are today, not where you hope to be in two years. Migrating between gateways is annoying but not fatal — most speak OpenAI-compatible APIs, so the switching cost is real but bounded.

Generated 2026-08-01 · Read the live article at scored-tools.netlify.app/blog/best-ai-api-gateway-tools-llm-applications-2026