ROUNDUP

Best AI Gateways for LLM Inference Optimization 2025–2026

Seven AI gateways compared on caching, routing and real fees, with 2026's acquisitions and the LiteLLM compromises factored in.

Introduction

An AI gateway sits between your app and the model providers. It leaves the model's own speed alone, so the savings come from around the call. A gateway can serve repeat requests from a cache, keep a conversation on the provider that holds its prompt cache, move traffic when a provider fails, and stop spend at a budget.

The vendor picture changed a lot in 2026, and a list written in 2025 is now out of date. Palo Alto Networks closed its purchase of Portkey on May 29, 2026. Stripe agreed to buy OpenRouter on August 19, 2026, and the latest reporting we found still shows that deal pending. LiteLLM had two PyPI supply chain compromises in seven months.

Three kinds of caching get lumped together. Exact-match response caching returns a stored answer for an identical request. Semantic caching returns a stored answer for a similar request, which needs embeddings and a vector store. Prompt caching happens at the provider and discounts a repeated prefix. Vercel's docs put Anthropic's rates at 1.25× base input for a five-minute cache write and 0.1× for a read, so one follow-up read pays for the write. LiteLLM's docs warn that semantic caching suits single-shot prompts and 'goes badly wrong on agentic traffic', so agent builders should start with prompt caching.

scored.tools does not test tools hands-on. Everything below comes from vendor pricing pages, vendor docs and named reporting, read in November 2026. Our scores come from five criteria: quality of output, ease of use, pricing value, API and integration quality, and how well a tool solves the problem it claims to solve. No rating data for these seven tools was available for this piece, so no entry carries a score. The order is alphabetical and implies no rank.

The gateways

Bifrost

Bifrost is an Apache 2.0 gateway from Maxim that puts one OpenAI-compatible API over 23+ providers. Its README reports 11 µs of added overhead per request at a sustained 5,000 RPS on a t3.xlarge instance. TrueFoundry, a competitor, points out that this is a synthetic load on one instance type and excludes provider latency, and the README's own About section uses a looser 'under 200µs'. Caching is the stronger reason to look at it. The docs describe a direct hash match first, then a semantic match at a default similarity threshold of 0.8. Setup takes work, because a vector store is required even for direct matching and requests without a cache key bypass the cache. The caching docs do not say whether the feature sits in the open source or enterprise edition, so confirm that before you plan around it.

- **Score** not yet rated
- **Best for** teams that want a self-hosted gateway with semantic caching and can run a vector store.
- **Pricing** is free for the open source edition on your own infrastructure. Enterprise is custom-priced, and Maxim advertises a 14–day trial.

Cloudflare AI Gateway

Cloudflare Ai Gateway has the lowest barrier here, since analytics, caching and rate limiting are free on all plans. Its cache is exact-match only, and the docs list semantic search as a future plan. Per-request cache TTLs run from a minimum of 60 seconds to a maximum of one month. Dynamic routing adds more for cost control than the cache does. Percentage nodes split traffic for gradual rollouts, while rate limit and budget limit nodes send traffic down a fallback path once a key hits its cap. The bill gets harder to predict as you add extras. Gateways created on or after September 24, 2026 store logs under Workers Logs pricing, and Guardrails bill as Workers AI inference, so that cost grows with prompt and response length.

- **Score** not yet rated
- **Best for** teams that want exact-match caching and per-key budget caps with no new infrastructure.
- **Pricing** is free for core features. Unified Billing adds a 5% fee on credit purchases, and logs, Logpush, Guardrails and full DLP profiles are priced separately.

Kong AI Gateway

Kong Ai Gateway suits organizations that already run Kong for APIs, and the published rates make that plain. A serverless gateway on Konnect Plus costs $25/month per control plane, but each unique LLM costs $100/month, with a limit of 5. By our arithmetic from those rates, five models on one serverless control plane come to $525/month before overage. Plus includes 1 million requests a month, and each further million costs $200, up to 10 million. AI Semantic Cache is listed as an enterprise plugin, included on Enterprise and offered as an add-on with no listed price on Plus. A small team pays a lot here for what Cloudflare or Vercel charge little or nothing for, so the case rests on policy and governance in a stack you already operate.

- **Score** not yet rated
- **Best for** platform teams already on Kong that want AI traffic under the same controls as their APIs.
- **Pricing** starts from $25/month plus usage on Plus. Hybrid gateways cost $200/month per control plane, and Enterprise is custom.

LiteLLM

Litellm offers the widest caching choice of the self-hosted options, and it carries the worst security record of 2026 in this group. It describes itself as an open source gateway for 100+ LLMs, and its README cites 8ms P95 latency at 1k RPS. The proxy supports exact-match caching on Redis, Valkey, S3, GCS, memory and disk, plus semantic caching on Redis, Valkey and Qdrant. The problem is supply chain history. Backdoored releases 1.82.7 and 1.82.8 reached PyPI on March 24, 2026, and a poisoned 1.107.0 reached PyPI and Docker on October 21, 2026. Help Net Security reports that the second attack targeted cloud logins and AI service API keys, and that Enterprise customers on the private registry were unaffected. A gateway holds every provider key you own, so this weighs heavily. Pin versions, and treat credentials as compromised on any host that installed those builds.

- **Score** not yet rated
- **Best for** Python-heavy teams already on it, with pinned versions and a strict key rotation process.
- **Pricing** is free for the open source version, plus your infrastructure. Enterprise is custom-priced with a 30-day trial key, and adds SSO, audit logs and air-gapped deployment.

OpenRouter

Openrouter is the quickest route to model breadth, with 600+ models across 80+ providers. It has the most complete hosted caching story in this group. Response caching is off by default and turned on per request or preset. Cache hits are free and skip provider rate limits, with a default TTL of 300 seconds and a range of 1 to 86400 seconds. The cache is scoped to one API key and is unavailable when account-level Zero Data Retention is enforced. Sticky routing sends later requests to the same provider endpoint to raise prompt cache hits, with sessions expiring after 10 minutes of inactivity. The fee is the trade-off, at 5.5% of credit purchases on Standard and 8% on Business. Stripe will own the company once the deal closes, and OpenRouter says its product and model-neutral approach stay as they are.

- **Score** not yet rated
- **Best for** teams that switch models often and want routing, response caching and prompt cache stickiness in one hosted service.
- **Pricing** has a free plan capped at 50 requests a day, rising to 1,000 once you have bought $10 of credits. BYOK includes $25,000 of list-price inference a month with no fee, then 5% of list-price cost.

Portkey

Portkey packs a full set of gateway controls into a low published entry price. Fallbacks, load balancing and retries appear on every plan, and the docs say simple caching does too, with a default TTL of 7 days and a maximum of 90 days. Semantic caching is limited to select Enterprise plans and needs a vector database. The MIT-licensed gateway on GitHub claims under 1ms latency and a 122kb footprint, and semantic caching is marked as hosted and enterprise only. The free Developer plan records 10k logs a month with 3 days of retention, and Portkey itself calls it 'Not suitable for production workloads'. Palo Alto Networks plans to integrate Portkey into Prisma AIRS, so check the roadmap and the self-serve plans before you sign anything long-term, especially if you are not a Palo Alto customer.

- **Score** not yet rated
- **Best for** teams that want managed fallbacks, caching and logs at a fixed monthly price.
- **Pricing** is free on Developer. Production is $49/month plus $9 per additional 100k requests, with 100k logs and 30 days of retention. Enterprise is custom.

Vercel AI Gateway

Vercel Ai Gateway charges no markup on tokens, including with your own keys, which makes it the cleanest fee structure among the hosted resellers. It works from any infrastructure, so you do not need to deploy on Vercel. For agent workloads, setting caching to auto adds cache markers for Anthropic, MiniMax and Alibaba, and session affinity helps keep a conversation on the provider that holds its cache. The caching docs cover provider prompt caching and describe no gateway-level response cache, and Vercel says prompt and response content is deleted once a request completes. Two BYOK details deserve attention. Budgets exclude BYOK spend, and a failed BYOK request can fall back to system credentials charged to your balance, so set provider-side limits as well. Some controls also cost extra, including team-wide Zero Data Retention at $0.10 per 1,000 requests.

- **Score** not yet rated
- **Best for** agent and multi-turn apps where prompt caching and provider failover matter more than repeat-request caching.
- **Pricing** is provider list price with no platform fee. A monthly free credit covers a subset of models, and BYOK requires purchased credits, which expire one year after purchase. Payment processing fees are yours.

Left out: Helicone

Helicone still publishes plans from free to $799/month, and its pricing page lists caching, rate limits and automatic fallbacks. Mintlify bought it in March 2026, and reviewers including ChatForest describe it as in maintenance mode. Its site now reads 'Helicone is no longer part of Mintlify. Details to come soon.' We would not start a new deployment on it until those details arrive.

Comparison table

Tool	Runs as	Response cache	Semantic cache	Entry price	Main fee to watch
Bifrost	Self-hosted	Yes, needs a vector store	Yes, edition unclear in docs	Free open source	Your infrastructure
Cloudflare AI Gateway	Hosted	Exact match	No, listed as a future plan	Free core features	5% on Unified Billing credits, plus logs
Kong AI Gateway	Konnect serverless, hybrid or dedicated	Not confirmed in the pricing reviewed	Enterprise, or add-on on Plus	From $25/month plus usage	$100/month per LLM on Plus
LiteLLM	Self-hosted	Exact match, several backends	Redis, Valkey, Qdrant	Free open source	Your infrastructure and security work
OpenRouter	Hosted	Exact match, opt-in	Not described in the docs reviewed	Free plan, 50 requests a day	5.5% or 8% on credit purchases
Portkey	Hosted or open source	Simple caching	Select Enterprise plans	$49/month for production use	$9 per additional 100k requests
Vercel AI Gateway	Hosted	Not described in the docs reviewed	Not described in the docs reviewed	Monthly free credit	Per-request add-ons

Final picks

These picks follow from published pricing and docs. They are not score rankings, since this roundup has no scores.

- **Lowest-cost start:** Cloudflare AI Gateway. Exact-match caching and rate limiting are free, and budget nodes cap spend per key.
- **Agents and multi-turn apps:** Vercel AI Gateway. It adds no token markup, and automatic cache markers go after the prompt caching discount that matters most for this traffic.
- **Frequent model switching:** OpenRouter. The catalog is the largest here and free cache hits help repeat traffic, though you pay 5.5% on credits and take on an ownership change.

- **Managed controls at a fixed price:** Portkey Production at $49/month, provided the Palo Alto roadmap suits you.
- **Self-hosted:** Bifrost for new deployments, given LiteLLM's two compromises in 2026. Existing LiteLLM users should pin versions and audit for the affected builds before they plan any migration.
- **Existing Kong shops:** Kong AI Gateway makes sense inside a stack you already run. The per-model pricing is hard to justify otherwise.

Generated 2026-11-06 · Read the live article at scored.tools/blog/best-ai-gateways-llm-inference-optimization-2025-2026/