Gateway	Models supported	Routing intelligence	Observability depth	Pricing	Best for
Inworld Router	220+ models from leading providers through a single endpoint	Business-metric routing on cost, latency, engagement, or revenue, with CEL-based conditional rules on request metadata	Per-request logging of model selected, attempt chain, TTFT, token cost, and routing reason, with export to Mixpanel/BigQuery	LLM usage billed at provider cost with no markup, across all plan tiers	Cost-quality optimization at scale
OpenRouter	400+ models from 70+ providers	OpenRouter's default behavior is to load balance requests across providers, prioritizing price	Request-level logging and cost dashboards	No markup on per-token pricing; 5.5% platform fee on Pay-as-you-go credit purchases	Rapid prototyping across many models
LiteLLM	100+ model providers via unified API	Fallback chains and load balancing configurable through code	Request logging and spend tracking	Open-source; self-hosted at no licence cost, with production deployment requiring at least 4 CPU cores and 8 GB RAM plus a PostgreSQL database	Full infrastructure control for teams with engineering capacity
Portkey	3,000+ models and providers	Guardrail-based and fallback routing	Detailed logs, traces, cost analytics, and guardrail violation tracking	Developer tier free; Production $49/month; Enterprise custom	Compliance-driven enterprise deployments
Braintrust Gateway	Major providers including OpenAI, Anthropic, Google, and AWS via a unified API	Automatic caching and multi-provider routing through a single endpoint	Structured trace logging tied to model, inputs, outputs, timing, metadata, and cache status, with scores attachable after the fact	Starter plan free (1 GB processed data, 10K scores/month); Pro $249/month; Enterprise custom with on-prem or hosted deployment	Teams that need detailed observability tightly coupled with output scoring
Helicone	100+ models via proxy, including OpenAI, Anthropic, and others	Routing relies on automatic failover across providers rather than conditional logic based on request content or business rules	Cost dashboards, per-request logging, and usage analytics	Open-source (Apache 2.0); free tier of 10,000 requests/month	Startups and small teams wanting a low-friction analytics layer over LLM calls