DropTicks AI Article

LLM Gateways and Model Routing Are Becoming Core AI Infrastructure

LLM Infrastructure | Mar 12, 2026

Why teams need model routing, budgets, fallback logic, observability, and provider abstraction before AI usage scales.

Article

When a team has one prototype, it can call one model directly. When a team has multiple workflows, customers, cost limits, and reliability expectations, direct model calls become messy.

An LLM gateway gives the organization one place to manage model access. The useful features are:

1. Unified API shape across providers.
2. Per-team or per-user budgets.
3. Fallback routing when a provider fails.
4. Cost tracking.
5. Logging and observability.
6. Guardrails and content filters.
7. Key management and access control.

This does not mean every small app needs a gateway on day one. But a company building multiple AI systems should plan for one. It prevents each product from inventing its own provider wrapper, retry rules, and cost tracking.

The practical rule: if a business has more than one production AI workflow, centralize model access before it becomes expensive to unwind.

Web link

https://www.dropticks.com/articles/llm-gateways-and-model-routing-are-becoming-core-ai-infrastructure

Related links

https://docs.litellm.ai/docs/
https://openai.github.io/openai-agents-python/
https://arize.com/docs/phoenix