DROPSOLID
ARTIFICIAL INTELLIGENCE

GUIDE 04

TECHNICAL GUIDE: THE MODEL GATEWAY DECISION FRAMEWORK

Choosing your LLM

A decision framework for the model gateway.

Published by

Dropsolid AI · 2026

WHY THIS GUIDE EXISTS

The right choice today isn't the right choice in twelve months

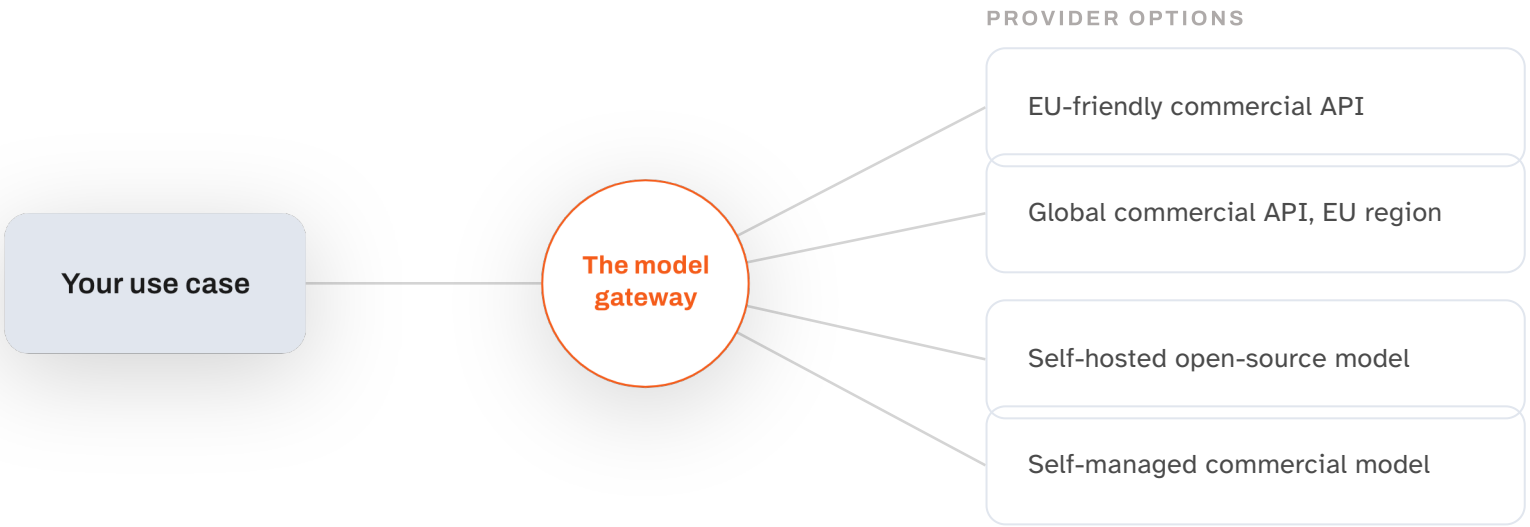
You're going to choose more than one language model over the lifetime of your AI deployment. The right choice today is not the right choice in twelve months. Model quality, cost, sovereignty options, and licensing all change quarterly.

The model gateway is the architectural pattern that lets you choose, swap or self-host each model per deployment without rebuilding the use case. This guide is how to think about the decision.

LLM
The language model that generates the answer.

Model gateway
The layer in between that routes each request, so models and providers can be swapped.

Provider
The company or infrastructure that runs that model.



THE FIVE VARIABLES

What actually narrows the model set

VARIABLE 1

Sovereignty requirement

Where must model inference happen? Full EU-resident, EU-resident-with-audit, self-hosted, on-premise.

VARIABLE 2

Governance requirements

What audit access, DPA position, and subprocessor disclosure does the model provider offer? For public-sector deployments, this often eliminates the majority of providers before latency or cost enter the conversation.

VARIABLE 3

Language coverage

Which languages must the model handle with source-grounded output? All 24 EU languages, a regional subset, or specialist languages (Basque, Welsh, Luxembourgish)?

VARIABLE 4

Latency budget

Public chatbots need answers under two seconds. Internal knowledge search can tolerate five. Some use cases are asynchronous.

VARIABLE 5

Cost per interaction

Not per API call. Per full user interaction, including retrieval, embedding, and inference. Costs can vary by an order of magnitude across providers for the same task.

Each level narrows the model set.

Ordered by typical filtering power.

PROVIDER CATEGORIES

Four provider categories, four trade-offs

Commercial EU-friendly APIs

European-based providers offering GDPR-aligned inference. Sometimes the fastest path to a viable production deployment when sovereignty is important but not extreme.

Commercial global APIs with EU regions

Major US-domiciled providers with EU-hosted API endpoints. Faster than European alternatives on raw model quality; require careful DPA review.

Self-hosted open-source

Full sovereignty. Higher operational complexity. Model quality is now competitive with commercial options for most tasks.

Self-managed commercial models

Some commercial providers allow you to run their model on your own infrastructure under specific licensing. Middle ground. Not always cheaper.

NO CATEGORY IS UNIVERSALLY BEST.

Which one fits falls out of the five variables — and with a gateway in place, the choice isn't final. The matrix on the next page turns it into a working brief.

THE DECISION MATRIX

Fill this in for your deployment

Turn the five variables into a working brief: one row per variable, your requirement, and the constraint it puts on providers. Examples shown in gray.

VARIABLE	YOUR REQUIREMENT	PROVIDER CONSTRAINT
1 **Sovereignty**	*e.g. EU-resident only*	*Narrows to EU-hosted or self-hosted providers*
2 **Governance**		
↳ Audit access	*e.g. full prompt/response logging*	*Excludes providers without audit APIs*
↳ DPA position	*e.g. standard contractual clauses accepted*	*Narrows to providers with EU DPA templates*
3 **Language coverage**	*e.g. all 24 EU languages*	*Narrows to large multilingual model families*
4 **Latency budget**	*e.g. under 2 seconds for public chat*	*Rules out high-latency, batch-oriented models*
5 **Cost tolerance**	*e.g. under €0.02 per interaction*	*Rules out premium frontier models at scale*
Migration cost if we change later	*e.g. must be swappable within the gateway*	*Rules out proprietary fine-tuned lock-in*

Fill this in. Bring it to your scoping conversation with Dropsolid or any AI vendor. If a vendor won't give you their position on every row, they're not ready for enterprise deployment.

HOW THE MODEL GATEWAY HELPS

Configure the decision, don't rebuild it

The model gateway lets you configure model choice per deployment, per use case, and per interaction type. Same use case at customer A and customer B can run on different providers.

Practical implications

1. Start on the fastest-to-deploy option that satisfies sovereignty. Ship the use case.
2. Monitor quality, cost and latency in production against your KPI.
3. Swap providers when the metric points at a better option, or when the sovereignty requirement tightens.

pidpa

CASE STUDY

Pidpa's rollout is the case study: initial deployment on one vector database provider, migrated during rollout to a self-hosted alternative on Dropsolid-managed EU infrastructure when the sovereignty posture tightened. The use case itself did not change; only the gateway configuration did.

Continue the conversation

Have a talk with Dropsolid about the five variables, and see how the model gateway answers each one in practice.

Talk to Dropsolid AI

Whitepaper 01 — AI sovereignty & Procurement →

DROPSOLID
ARTIFICIAL INTELLIGENCE

Read Whitepaper 01, AI sovereignty: a CIO's procurement guide for the EU AI Act, for what changes in 2026 and what to ask your AI vendor.