All models.
One endpoint.
Your premises, your price.

Most teams pick one model and live with the trade-off. **Mixture of Models (MoM) puts every model behind a single OpenAI-compatible URL** — open source on your own hardware, frontier models when you need them — and routes each request on the priority you set.

3
PREMISES: LOCAL, HYBRID, SOVEREIGN

$0.00
PER TOKEN ON MODELS YOU OWN

4
ROUTING STRATEGIES

1
ENDPOINT FOR ALL OF IT

WHAT "AUTO" OPTIMIZES FOR

Balanced
`default`

Quality against cost, 50/50. The setting most teams never need to change.

Cheapest
`budget first`

Lowest price that still does the job. For volume: tagging, extraction, triage.

Quality
`benchmark first`

When being wrong costs more than tokens: customer-facing work.

Fastest
`latency first`

Throughput and time-to-first-token. For anything a person is waiting on.

Set it once per API key, or send it per request. **MoM reads what the call actually needs** — tools, vision, JSON mode — drops every model that cannot serve it, then ranks the rest by your strategy and holds the next-best as automatic fallback.

MAKE YOUR FIRST REQUEST

cURL | Python | Node.js

```
curl -X POST https://api.eggsplain.com/v1/chat/completions \
  -H "Authorization: Bearer sk-egg-..." \
  -H "Content-Type: application/json" \
  -d '{
    "model": "auto",            # MoM picks the model
    "strategy": "quality",       # or cheapest|fastest|balanced
    "messages": [{"role": "user",
      "content": "Explain sovereign AI in one sentence."}]
  }'
```

PYTHON SDK · OPENAI-COMPATIBLE

```
# same client, same call — only base_url changes
client = OpenAI(base_url="https://api.eggsplain.com/v1")
r = client.chat.completions.create(
    model="auto", messages=[…])
```

THREE PREMISES, ONE API

PREMISES	WHAT RUNS THERE	PER TOKEN
On-Premises	Open source & your own	`$0.00`
Hybrid-Premises	Local plus hosted	`mixed`
Sovereign-Premises	Opus, GPT, Gemini	`metered`

Own the card and the marginal cost of a token is **zero, forever**. Burst to frontier models when a job needs them and **pay per token only for those calls** — same endpoint, same strategies. **You set which premises each key may reach.**

Choosing the model should not cost a model call.

`MoM routing is a scheduling decision, not an inference one — it adds no thinking tokens and no metered spend to the request it routes.`