Adora AI OS ▸ Multi-Model Strategy & Cost-Control Framework

Investor-ready overview – May 2025

1. Executive Summary

Adora AI OS routes every task to **the least-expensive model that still meets quality SLAs**, escalating only when complexity or risk demands it. This granular control—down to each workflow step—cuts runtime LLM spend by **≈ 90 %** versus a single-model approach while preserving best-in-class performance for mission-critical steps.

2. Market Need & Gap

Most enterprise AI platforms lock clients into a "one-size, one-price" model tier, leaving cost and latency unpredictable. Enterprises need:

1. **Predictable unit economics** that scale with usage.
2. **Quality-of-service guarantees** for sensitive processes (healthcare, legal, finance).
3. **Governance hooks** to throttle or pause AI activity under budget or compliance constraints.

Adora AI OS meets these needs by combining fine-grained **Model Tiering** with **Time, Rate, and Token Governors** built into every agentic workflow.

3. Model Portfolio & Routing Logic

Tier	Default Model (Floor)	Escalation Model (Ceiling)	Typical Use Cases
T0 — Boilerplate	Llama-3 70B	n/a	Regex extraction, profanity check
T1 — Light Reasoning	Gemini Flash-Lite / GPT-4o-mini	Claude 3.5 Haiku	Summaries, email drafts

T2 — Moderate Reasoning	Mistral Medium 3	GPT-4o-mini	RAG answers, light code
T3 — Heavy / Vision	GPT-4o-mini	GPT-4o	Code gen, screenshot UI agenting
T4 — Edge-case Excellence	GPT-4o	o3 / Claude 3 Opus	Long-context legal, medical

➡ Routing Engine

- Lightweight classifier (Llama-3 8B) scores each request for *complexity, context-length, risk*.
- Maps score → Tier, sends to floor model.
- Automated tests & hallucination heuristics retry → escalate if needed.

(Interactive price table provided separately.)

4. Built-in Cost & Latency Governors

Governor	Scope	How It Works	Business Value
Time Cap	Step or full workflow	Abort or downgrade if step exceeds X seconds.	Prevent runaway latency; meet SLA.
Rate Limit	Org / User / Model	Max requests per minute; queue overflow or invoke cheaper model.	Smooth spikes; avoid burst overage fees.
Token Cap	Step or full workflow	Hard ceiling on tokens used; auto-summarise or split tasks.	Predictable budget per job.
Delay Window	Batchable tasks	Re-route to Batch API (-50 % cost) if task not time-critical.	Turns overnight slack into savings.
Prompt Cache	Global	48 h cache; shared prefixes charged at 50 %.	~30 % extra savings on repetitive ops.

Governors are **config-as-code** so Finance & Ops can tweak limits without redeploying agents.

5. Sample Workflow Walk-Through

Invoice → Payment Reconciliation (8 steps | 3 models | 2 governors)

1. **OCR & Parse PDF** · *T0* · Llama-3 70B
2. **Entity Extraction** · *T1* · Gemini Flash-Lite · *Token Cap 5 k*
3. **Line-Item Validation** · *T2* · Mistral Medium 3
4. **Exception Check** · (T1) If fails → *T3* GPT-4o-mini
5. **Ledger Entry Draft** · *T1* · Gemini Flash-Lite
6. **Compliance Review** · *T3* · GPT-4o-mini · *Time Cap 3 s*
7. **Audit Log Embedding** · *T0* · Edge Llama-3
8. **Supervisor Summary** · *T2* · Mistral Medium 3 → triggers email via Function Calling.

Outcome: Avg cost $0.0012/invoice, 1.9 s P95 latency, full rollback if Token Cap breached.

6. Competitive Advantage

- **Granular Controls** – Time/Rate/Token governors at *per-step* granularity (competitors apply at tenant level at best).
- **Auditing Agent & Dual-Model Checks** – automatic validation in high-risk flows.
- **Edge Off-load** – Micro-data-centers run OSS models at <$0.05/M tokens power cost.

7. Financial Impact Snapshot

Launch cohort (1 000 orgs × 20 seats):

- **Raw GPT-4o Only:** $13.5 M/month LLM bill
- **Adora Tiered + Governors:≈ $0.9 M/month**
- **Net Savings:** 93 % ↓ (see cost-stack spreadsheet)

8. Roadmap & Next Milestones

Qtr	Milestone	Impact
Q2 ’25	Benchmark harness live	Data-driven routing tuning
Q3 ’25	Governance UI for clients	Self-service spend caps

Q3 ’25	Enterprise commits negotiation	-12 % token unit cost
Q4 ’26	First micro-data-center online	Divert 20 % traffic to edge