THE BUILDER'S STACK · THE FRONTIER · THE UNIT ECONOMICS OF AI

The Inference Budget

One budget that makes inference an economic decision: the unit of work, the speed it owes the user, the models and caches that serve it, and the price that protects the margin. Fill it before launch; reopen it the day a review trigger fires.

PRODUCT / FEATURE

WRITTEN BY

1 The unit

The task, defined tightly enough to cost: what one run costs today, and the number it must reach.

ONE UNIT OF WORK IS (THE TASK, DEFINED)

PER-TASK COST TODAY

THE TARGET IT MUST REACH

2 The latency budget

What the user is owed per feature class: when the first token lands, and when the answer finishes.

FEATURE CLASS	FIRST TOKEN BY	COMPLETE BY

3 The routing plan

Which traffic gets which model, what sends a request up a tier, and the eval bar each model cleared first.

TRAFFIC	MODEL	EVAL BAR THAT AUTHORIZED IT

THE ESCALATION GATE (WHAT SENDS A REQUEST TO THE STRONGER MODEL)

4 The caching plan

Check only the caches this workload actually earns; every checked cache gets a lifetime.

CACHE	LIFETIME
☐ **Prompt cache** — the shared prefix, reused across calls	
☐ **Response cache** — the exact same question, answered once	
☐ **Semantic cache** — close-enough questions, one stored answer	

5 The realtime split

Nobody pays realtime prices for work nobody is watching: sort every job into a lane.

RUNS REALTIME (THE USER IS WATCHING THE ANSWER ARRIVE)

RUNS IN THE BACKGROUND (THE USER GETS THE RESULT WHEN IT LANDS)

RUNS AS A BATCH (NOBODY IS WAITING)

6 The pricing stance

The currency you charge in, what the plan includes, and the margin where the users actually sit.

- ☐ **Seat** — a flat rate per person
- ☐ **Usage** — metered by consumption
- ☐ **Outcome** — paid on the result
- ☐ **Hybrid** — a base plus a meter

THE INCLUDED ALLOWANCE (WHAT THE PLAN COVERS BEFORE OVERAGE)

MARGIN AT THE MEDIAN USER

MARGIN AT THE HEAVY USER

ALARMS

Each alarm gets a threshold that fires before the invoice does, and one person who answers it.

ALARM	THRESHOLD	OWNER
Per-task cost		
Escalation rate		
Cache hit rate		

SIGN-OFF

This budget reopens on a trigger, not a calendar: a provider reprice, or a shift in the workload.

ACCOUNTABLE OWNER

DATE

REVIEW TRIGGERS

PRESSURE TEST

Walk each scenario against the filled budget. Check the box only when the budget already answers it.

- ☐ Adoption triples: the per-task cost holds, the alarms stay quiet, and the margin at the heavy user survives.
- ☐ The provider halves prices: the routing plan says what moves down a tier, and the pricing stance says who keeps the savings.
- ☐ One customer goes heavy overnight: the allowance absorbs it or the meter starts, the escalation alarm fires before the invoice does, and the margin still closes.