ARCHITECTURE NOTES · RESOURCE

Per-token costs are trivial. Per-outcome costs are everything.

Most teams cannot compute the cost that matters because the costs that matter live outside the model bill. Four levers fix the math.

Companion to: Per-token costs are trivial. Per-outcome costs are everything.

Why the model bill is the wrong number to optimise

A short social post drafted by a current-generation model costs a fraction of a cent. The same post by a freelance copywriter costs hours. The gap is roughly three orders of magnitude, which is why every AI-content vendor cites it on slide three. The number that matters is what it costs to ship a post that performs at parity with a human-written one. That bill includes operator review minutes, the iterations to fix register, the drafts thrown out, and the off-brand outputs that ship anyway and quietly erode trust.

- Where per-outcome costs actually live
- The same shape outside content (retention, CRO)
- Four levers that change the math
- Where the math flips and the structural work does not pay back

The four levers that change the math

01 **Brief schema**

The model needs the same input a human writer would need: pillar, voice, hook, key points, CTA, source. Most teams give models a thinner brief than they would give a junior copywriter and then blame the model when copy comes back shallow. The result is an input problem

02 **Voice contract**

A written specification of what your brand sounds like, enforced as a system prompt going in and a scanner running against every output. Drafts that fail the scanner get rewritten before any human sees them. This layer is what prevents the slow brand drift that sinks most

03 **Observability**

Logs structured around outcomes. When a draft is bad, you need to know in seconds whether the brief was thin, whether the voice gate caught it, or whether the model regressed. Generic LLM logging tells you which API call returned what; it does not tell you which brand

04 **Iteration discipline**

Programs that succeed iterate the brief. Programs that fail iterate the model: switching providers, escalating to a larger one, commissioning custom fine-tunes, and never returning to the brief that produced the bad draft in the first place. Get the lever right and a competent model is

Per-token vs. per-outcome

Per-token (the cheap number)

- Fraction of a cent per social post draft
- Cited on slide three of every AI vendor deck
- Ignores review minutes, retries, and drafts thrown out
- Says nothing about parity quality
- Same shape on Klaviyo and Webflow: the tooling line is cheap

Per-outcome (the real number)

- Cost to ship a post that performs at parity with a human-written one
- Includes operator review, iterations, register fixes, off-brand drift
- Includes the architectural work the cheap version skipped
- On retention: deliverability hygiene the template version did not earn
- On CRO: the architectural work of a real build, regardless of who wrote draft one

The same shape outside content

01 **Retention: cheap CPM vs. real conversion**

A Klaviyo welcome flow billed by sender CPM is cheap. A welcome flow that actually converts at parity with a hand-built one costs the same architectural work the hand-built one would, plus the deliverability hygiene the cheap version skipped. The model bill is not where the cost

02 **CRO: free template vs. real Lighthouse build**

A Webflow page generated from a template is free in tooling. A Webflow page that scores 100 on Lighthouse and earns its place in the funnel costs the architectural work of a real CRO build, regardless of whether a model wrote the first draft.

03 **Where the math flips**

A content drafter for a brand publishing once a week, a retention agent for a small list, a CRO test stack for a page receiving a trickle of traffic. None of these exercise the contract layer often enough to amortise it. Below the threshold, the structural work does not pay back. This is why

Published program-level outcomes

AI LAB PROGRAM-LEVEL LIFT

20% retention

Aggregate retention lift across operations we have rebuilt

OPERATOR TIME SAVED

60%

Program-level time-saved aggregate, AI Lab

YOY REVENUE

Doubled

Across operations rebuilt under the program

Operator runbook: compute the per-outcome number

- **Define parity**
  What does a human-written version of this output cost in time, dollars, and quality? Without that anchor, no per-outcome number is meaningful.
- **Sum the hidden costs**
  Operator review minutes, retries, drafts thrown out, off-brand outputs that shipped and quietly eroded trust. The model bill is a rounding error against this stack.
- **Audit the brief schema**
  If your model brief is thinner than a junior copywriter brief, the input problem is wearing a model-problem costume. Fix the brief before changing providers.
- **Install the voice contract before scaling volume**
  A scanner that catches off-brand drafts before any human sees them. Without this layer, programs drift by month three regardless of how good the model is.
- **Structure logs around outcomes, not API calls**
  You need to know in seconds whether the brief was thin, the voice gate caught it, or the model regressed. Generic API logging cannot answer that.
- **Iterate the brief, not the model**
  When a draft is bad, the first move is to fix the input that produced it. Switching to a larger model is the move teams make instead of doing the work.
- **Confirm scale fit before signing**
  Below the threshold where the contract layer is exercised often enough to amortise, structural work does not pay back. Recognise that and walk.

When the per-outcome math fails

01 **Below the engagement threshold**

A retention agent on a small list or a CRO stack on trickle traffic does not exercise the architecture often enough to amortise it. The published threshold on /websites-cro is 30 to 50 thousand euro per month in brand revenue, on the page on purpose.

02 **Optimising the wrong line**

Teams that swap providers, escalate model size, or commission fine-tunes before fixing the brief are paying premium per-token rates to mask a per-outcome problem. The model bill goes up; the per-outcome cost stays bad.

03 **No voice scanner**

Without a voice contract enforced before any human sees the draft, the brand drift compounds quietly until month three. By then the per-outcome cost includes the cost of pulling the program back from drift.

04 **Logs that report calls, not outcomes**

When you cannot tell in seconds which brand decision broke, every bad draft costs a long human investigation. Generic LLM logging is one of the largest hidden lines in per-outcome economics.

NEXT STEP

Optimise the per-outcome number

A competent model is enough when the four levers are right. The wrong levers turn the most expensive vendor into middling output.

Read the full architecture note ->