Cut Your AI Bill · short summary · built 11 Nov 2026 ·
demo company: Corvid Cloud

Corvid Cloud support AI: where the LLM bill goes

This is the short summary of the report you receive from Cut Your AI Bill. Every token count, cost and quality result below comes from a real run of the pipelines described in the method section. The full report adds the code change for each fix, risks and a 30-day verification plan.

$23,117 → $531 a month

97.7% lower at the list prices of the models we ran · quality: 76 of 81 checks pass (78 before)

---

1. Ranked fix list

Levers were applied in order on the same workload; each saving is measured on what the previous levers left. Ordered here by monthly dollars.

#	FIX	SAVES / MONTH	OF THE BILL
1	**Trim the prompt and retrieved context** A shorter system prompt with every rule still in it, and the 3 best-matching knowledge-base passages instead of 10.	$13,410	58.0%
2	**Cap output and reasoning** Shorter answer formats, a length limit per task, and less hidden reasoning on the big model.	$5,000	21.6%
3	**Route easy calls to a cheaper model** Classifier, summaries and nightly digests go to the small model; routine support questions too. Doubtful answers are re-run on the big model.	$3,488	15.1%
4	**Prompt caching** The part of every prompt that never changes (policies, label definitions) is billed at the provider's cheaper cached rate.	$639	2.8%
5	**Dedupe repeated calls** A repeated or reworded question is answered from a response cache, without a model call.	$26	0.1%
6	**Batch the nightly job** The nightly digest is not urgent, so it runs as a batch job at half price.	$23	0.1%

Monthly figures assume 1,200,000 requests a month in the same mix as the 81-request sample. Each call is priced at the list price of the model that served it.

2. Price comparison: the same tokens on each provider’s list

The result above prices each call at the model that served it, the As run row. For comparison only, the other rows price the same measured tokens on each provider’s list.

PRICE LIST	BIG / SMALL TIER PRICED	BEFORE	AFTER	SAVED
As run (the result)	Claude Sonnet 5.5 / Gemini 3.1 Flash-Lite	$23,117	$531	97.7%
Anthropic	Claude Sonnet 5.5 / Claude Haiku 5.5	$23,117	$433	98.1%
OpenAI	GPT-6.1 Sol / GPT-6 Luna	$23,117	$425	98.2%
Google	Gemini 4 Carbon / Gemini 3.1 Flash-Lite	$23,117	$524	97.7%

Comparison only. The result is the As run row.

Prices: public list prices read on 11 Nov 2026. Claude Sonnet 5.5, GPT-6.1 Sol, Gemini 4 Carbon all list at the same $2.00 in and $10.00 out per million tokens, so the before bill is the same on every list; the after bill moves with each small model's price.

3. Quality: checks passed

TASK	BASELINE	OPTIMIZED, FIRST PASS	OPTIMIZED, FINAL
Support answers	36/37	28/37	35/37
Ticket classifier	23/24	23/24	23/24
Ticket summaries	12/12	7/12	10/12
Nightly digest	7/8	1/8	8/8
All checks	**78/81**	**59/81**	**76/81**

Exact match on the correct label for the classifier; a grading model (Gemini 4 Argon, a second AI model) with a written rubric (correct, complete, policy-safe) for everything else. The escalation rule re-runs a cheap answer on the big model if the answer says it is unsure, cites nothing, or cites a passage that was not retrieved; arithmetic questions route straight to the big model. It re-

ran 4 answers. Two more fixes came from the quality check: digest counts computed in code and passed to the model, and a less terse cap (90 words) for summaries and support answers.

Checks that still fail in the final pipeline

- **R019** (summarize): Ticket T-7108 | Customer: Hiro (Business) Hiro: We need our data in the EU region for a new German client. Can you move our workspace? Agent
  baseline pass, first pass pass, final fail | grader's note on the final answer: Missed specifying EU region and creating a new EU workspace.
- **R031** (qa): My invoice has the wrong company name. Can you reissue last month's?
  baseline pass, first pass fail, final fail | grader's note on the final answer: Did not offer to check with the Billing team about reissuing last month's invoice.
- **R045** (qa): How do I set up a webhook and verify it's really from you?
  baseline fail, first pass fail, final fail | grader's note on the final answer: Omitted Settings > Integrations > Webhooks and falsely claimed setup location wasn't in the policy.
- **R049** (summarize): Ticket T-7109 | Customer: Dana (Team) Dana: Charged $1,740 on the 1st but we only have 60 seats at $29. Should be $1,740... wait, that's rig
  baseline pass, first pass fail, final fail | grader's note on the final answer: Missing the 'no action' part of the status.
- **R056** (classify): Is there a way to switch the Kanban board to dark mode?
  baseline fail, first pass fail, final fail

Shipping rule: a change that fails the quality check does not ship. The failing checks above are the open items for the next iteration, not hidden.

---

4. Method and assumptions

- **Workload.** 81 requests: 37 support questions answered from a 32-passage knowledge base, 24 ticket classifications, 12 ticket summaries, 8 nightly account digests.

- **Baseline.** One big model (Claude Sonnet 5.5) for everything, long system prompt, 10 retrieved passages, no caching, uncapped output, nightly job synchronous.
- **Optimized.** Easy calls go to a small model (Gemini 3.1 Flash-Lite): classification, summaries, digests and routine questions; a trimmed prompt with a stable prefix; the top 3 re-ranked passages; short answers with a length limit and low reasoning effort; a response cache for repeated questions (3 exact and 5 reworded repeats in 81 requests); the nightly digest priced at the batch rate.
- **Tokens.** Taken from the API usage field of each call. Stable-prefix sizes measured per model.
- **Caching.** Assumed 95% of the stable prefix is served from cache at scale, and only prefixes of at least 1,024 tokens are cached. Cache expiry and storage fees are reconciled against your invoice when we run this on your bill.
- **How we measured.** One workload, one run per stage, temperature 0, one grading model. On your traffic we build the quality test from your own requests and hold back 20% as a test set.

Get this report on your own bill

Send your AI bill export. The free AI check, a one-page report within 48 hours, tells you which levers apply and a dollar range, before you pay anything. Or check your bill free in your browser with the LLM Bill Checker. Get a free AI check: khlabs.dev/#check · hello@khlabs.dev

Or write to hello@khlabs.dev.