Your AI Bill Just Dropped Two-Thirds — Same Output

Perplexity just shipped a model that matches Claude Opus 4.8 at about a third of the cost — here are three ways to bank the savings this week without changing a thing you already built.

Source: SquaredTech

THE NEWS, IN ONE BREATH

On July 10, 2026, Perplexity released a fine-tuned GLM 5.2 model inside its Computer product that matches Claude Opus 4.8-level output at roughly one-third the token cost — about 34 cents for every dollar you were spending. Same tasks, same quality, a third of the bill. If you're running AI agents to handle admin, ops, or content, the price of the thing you already built just fell through the floor.

SO WHAT — IF YOU OWN THE BUSINESS

3 WAYS TO USE IT THIS WEEK

1 *Repoint what you already built — same workflow, a third of the cost*

15 MIN · OPS / COST · NO CODE

Do it: If your automation calls an AI model through a tool like Zapier, Make, or a no-code agent builder, most let you swap the model in a dropdown. Change it, run your usual task, compare the output. If it holds, you just cut that workflow's bill by two-thirds.

> **Where DIY stops:** Custom pipelines with hardcoded model calls, tuned prompts, or strict output formats need someone to swap and re-test them properly — that's where a repoint stops being a dropdown change and becomes a small build.

2 *A/B the cheaper model against your current one on real work*

30 MIN · OPS · NO CODE

Do it: Take five real tasks you run every week, push them through both models, and put the outputs side by side. Cheaper isn't a win if quality drops — this 30-minute test tells you whether the savings are real for your use case before you commit.

> **Where DIY stops:** Judging quality across hundreds of runs and edge cases needs a scored, repeatable eval harness rather than a manual eyeball check. That's a build, not an afternoon.

3 *Reinvest the saved two-thirds into the automation you kept putting off*

THIS WEEK · GROWTH · NO CODE

Do it: The workflow you skipped because 'the AI costs add up' just got three times cheaper. Re-price it. The follow-up sequence, the lead-list enrichment, the weekly report — the math that didn't work last month might work today.

> **Where DIY stops:** Stringing several automations into one unattended system — with error handling, approvals, and logging — is where DIY tools tap out and you're building infrastructure. That's the line where a Teardown pays for itself.

Here's what most operators miss: a price drop like this isn't a reason to wait for things to settle — it's a window. The businesses that win re-point their workflows the week the cheaper model ships, bank the savings, and put them straight back into the next automation. The ones that lose read the headline, think 'I'll look into it later,' and are still paying triple six months from now.

That's why ProofWright exists. ***I'm not an engineer either*** *— I built 10 of these for my own five companies, zero engineers hired. If you want to put this to work in your business — not as a weekend experiment, but as a system that runs while you're doing the actual job — book a Teardown. I'll map your process, show you the two or three automations that close your biggest time gap, and you'll leave with a plan you can execute. No deck, no theory — just the build.*

Which play fits your business right now?

Tell me what your business does — reply right here — and I'll map which of these three is cheapest to wire up for you now. No pitch, no deck.

Ready to build? Book a Teardown → **proofwright.ai/teardown** — 90 min, $500, credited toward any build in the next 30 days.

→ Reply with your industry. Worst case, you get a straight answer for free.