

AI TOKEN OPTIMIZATION

Every Token Should Finish Something.

Your AI bill measures activity, not results. Coefficient finds where token spend is leaking, fixes the routing and workflows behind it, and puts controls in place so every AI dollar ties back to finished work. We start with a five-day assessment, and we do not quote a savings number before we see your data.

4 months

How long it took Uber to use up its entire 2026 AI budget
CNBC, June 2026

90%

Discount on cached input tokens versus standard input price
Anthropic prompt caching pricing

50%

Discount for batch processing on work that does not need an instant answer
OpenAI and Anthropic batch APIs

5x

Input price gap between the largest and smallest current models from one vendor
Anthropic model pricing, 2026

Where Token Spend Leaks

Frontier models on routine work

Classification, extraction, and summarization sent to the most expensive model by default, when a smaller model does the job.

Context re-sent on every call

Long system prompts and reference documents paid for at full price on each request because caching was never turned on.

Real-time pricing for batch work

Overnight and back-office jobs run synchronously at full price when nobody needs the answer in seconds.

Retry loops and runaway agents

Failed calls and agent loops that burn tokens with no ceiling, no alert, and no one watching.

Orphaned jobs and seats

Scheduled jobs, abandoned pilots, and licenses that stopped paying off months ago and have no owner.

Rework nobody counts

Output that people quietly fix or redo. It never appears on the AI invoice, and it is usually the largest cost of all.

The Cost the Invoice Hides

ILLUSTRATIVE EXAMPLE • CLAIMS SUMMARIZATION

50,000 AI summaries at \$0.40 each	\$20,000 / mo
30% rewritten by adjusters, 10 min each at \$60/hr	\$150,000 / mo
Cost per summary on the invoice	\$0.40
Real cost per summary	\$3.40

Buy a model at half the price
Saves \$10,000. If rework rises to 35%, labor adds \$25,000. Net: \$15,000 a month worse, with a lower AI bill.

Fix prompt, context, validation
Rework falls to 10%. Saves \$100,000 a month. Real cost per summary drops to \$1.40. Same vendor.

Numbers are illustrative. The structure of the math is what matters.

OUR SCOREBOARD

Finished work per dollar.

A task is finished when the next person in the process accepts it without redoing it. The dollar includes tokens, tooling, and the people who check and correct the output. Cheap tokens on the wrong task are still waste. Some work needs a frontier model. Most does not.

THREE QUESTIONS FOR YOUR NEXT AI BUDGET REVIEW

- What did we spend, by workflow, and who owns each line?
- What share of the output was accepted without rework?
- What does a finished task cost now, compared with before AI?

If your team cannot answer all three, the spend is not managed yet.

THE LEVERS WORK

After its budget overrun, Uber reports it lowered its cost per token while expanding access, by improving prompt caching, changing its default model, testing newer models for efficiency, and showing engineers their own usage and cost. No new vendor required.

Fortune, August 2026

Sources: CNBC (June 26, 2026), "OpenAI and Anthropic face new AI spending reality as users shift to efficiency." Fortune (August 7, 2026), Uber CTO on AI cost per token. Anthropic prompt caching documentation (cache reads billed at 10% of base input price). OpenAI and Anthropic batch API pricing (50% discount). Anthropic model pricing, 2026 (Opus \$5 vs. Haiku \$1 per million input tokens). Worked example is illustrative, not a client result. Figures current as of September 2026; verify before external distribution where precision matters.

THE OFFER

The Five-Day Token Spend Assessment.

Fixed scope. Up to three priority workflows. Powered by the Coefficient Token Spend Monitor, which ingests your OpenAI and Anthropic usage exports and invoices, separates input, output, and cached usage, reconciles to billing, and flags missing data instead of inventing results.

- DAY 1

Set the baseline and agree on what "finished" means for each workflow.
- DAYS 2 TO 4

Analyze usage by provider, model, team, and workflow. Connect spend to finished work, rework, and quality. Rank the opportunities.
- DAY 5

Leadership readout with the six outputs below.

What You Receive

- 1

AI Spend Baseline

Spend by provider, model, team, and workflow, with billing reconciliation and every data gap called out.
- 2

Opportunity Register

Ranked opportunities with savings ranges only where the data supports them, plus effort, dependencies, and confidence.
- 3

Workflow Recommendations

Model choice, prompts, context, caching, batching, retries, and validation for each priority workflow.
- 4

Management Scorecard

Budget variance, cost per finished task, quality, response time, and rework, in one view leadership can track.
- 5

Spending Controls Playbook

Budgets, thresholds, accountable owners, escalation paths, and the action required at each threshold.
- 6

90-Day Roadmap

Executive readout with owners, sequencing, and measures of success for every recommended change.

Signs You Need This

- AI spend is up and operating results are flat.
- More than one team or provider is buying AI on its own.
- Nobody can name the cost per task for any AI workflow.
- The AI line ran over budget, or finance cannot explain it.

What We Need From You

- Provider usage exports and three to six months of invoices.
- Up to three priority workflows and a named owner for each.
- Time with the people who review AI output in those workflows.
- A leadership sponsor for the day-five readout.

Our Commitments

- No savings number before we see the data.
- Every opportunity comes with its evidence, method, and confidence.
- Where the data is incomplete, we say so.
- Quality is validated before any cheaper route is recommended.

After the Assessment

- 1

Assess

5 DAYS

Baseline, map, and ranked opportunities across up to three workflows.
- 2

Optimize

SCOPED FROM REGISTER

Implement routing, caching, batching, prompt and context redesign, and validation. Measure finished work per dollar before and after.
- 3

Control

ONGOING

Budgets, thresholds, named owners, and monitoring so spend does not drift back to volume for its own sake.

Built by Coefficient. Applied to your AI spend.

We have spent our careers finding the 10 to 45% in recoverable spend hiding in IT and BPO budgets, work that has generated \$500M+ in realized client savings across 12 Global Capability Centers. AI spend is a fast-growing line on that same budget. It responds to the same discipline.

START WITH FIVE DAYS

Want to know what your AI spend is actually finishing?

Contact Us →

coefficient-group.com
+1 724-344-6284 · linkedin.com/company/the-coefficient-group