OPTIMUM PARTNERS

AI Token Cost Checklist

10 questions to find avoidable AI spending and keep your budget under control.

Before you start

- Pick one AI workflow.
- Gather recent runs, including failures and costly cases.
- Use the same examples for every check. Savings count only if quality holds.

SCORE EACH CHECK — 0 Not checked, or no evidence · 1 Investigated, but a gap remains · 2 Verified with current data or a test

FIND WHERE THE BUDGET GOES

1	**Can you name an owner and a cost for every AI workflow?** *Bills and usage matched to apps and teams, including shared keys and scheduled jobs.* Where to look for untracked usage	Essential, Finance, IT	0	1	2
2	**Do you know what one successful task costs?** *Model, tool and human correction costs per batch, divided by successful completions.*	Finance, Ops	0	1	2
3	**Have you estimated spend at normal and peak volume?** *Measured task cost times expected volume, plus a busier scenario with harder cases.* Go deeper on workflow costs	Finance	0	1	2

REMOVE WORK YOU ARE PAYING FOR TWICE

4	**Does each step receive only the context it needs?** *Duplicate material and old history trimmed, with a test showing quality holds.*	Engineering	0	1	2
5	**Have you tested reusing repeated work?** *Prompt caching and skipping unchanged records, with net savings after write charges.* Explore the architectural sources of spending	Engineering	0	1	2

REDUCE THE COST OF COMPLETING THE TASK

6	**Is each task on the cheapest model and effort that works?** *Cheaper model and lower reasoning effort tested separately on routine and hard cases.* Explore model selection	Engineering	0	1	2
7	**Does the task need every step it takes?** *Calls and handoffs traced, a simpler sequence tested, fixed rules moved into code.* Understand why the surrounding system matters	Engineering, Ops	0	1	2
8	**Does the model receive the right evidence?** *Passages checked against sources for lost headings, split exceptions, old versions.* Improve the source material	Engineering, Data	0	1	2

KEEP SPENDING UNDER CONTROL

9	**Do failed runs stop, or reach a person, instead of looping?** *Recurring failures fixed, retry, step and time limits set, stuck cases tested.*	Essential, Engineering, Ops	0	1	2
10	**Does someone review costs, and do alerts reach a person who can act?** *Weekly review of cost per task, retries and correction time; alerts and pause tested.*	Essential, Ops, Finance	0	1	2

Your Total

____ / 20

0–7
Major gaps
Build reliable spending data and controls before expanding usage.

8–14
Partly controlled
Prioritise gaps behind the largest costs or repeat failures.

15–20
Well covered
Keep reviewing, and repeat checks when the workflow changes.

START HERE Fix your two lowest scores first. A zero on an Essential check needs attention whatever your total.

Want help reviewing these 10 architectural checks for your environment?

Reserve a zero-pitch, 20-minute token architecture review with Joe Sticca at Optimum Partners.

Book 20-Min Review →