NEXTUP SOCAL · AI WEBINAR SERIES

The AI Cost Checklist

A companion to *Reducing AI Operating Costs*. Everything here can be started without a budget request. Work top to bottom — the order matters more than the percentages.

Fri Aug 21, 2026
Nick McCarty
Upskilled Consulting

Four moves that don't need a budget

Start here. Two of these an executive can order without a technical conversation.

- [ ] **Meter it — you cannot manage what you're not measuring.** — This week
  Ask for cost per team, per month, and keep building your intuition for cost per workflow. Most platforms have this reporting switched off by default; turning it on is an afternoon of work.
  Done when: someone can tell you last month's cost for one named workflow.
- [ ] **Route it — name your three biggest use cases.** — 2 weeks
  Work through two of them using the cheaper model for two weeks and compare the output. If no material difference can be discerned, you've found savings.
  Done when: a written rule says which tasks use which tier.
- [ ] **Automate one thing.** — 1 month
  Pick the task that runs the same way every week — the Friday report nobody enjoys building. Ask for it to be automated once, and for the recipe to live somewhere the team can find it instead of one person's inbox.
  Done when: the recipe is in shared storage and someone else can run it.
- [ ] **Budget per outcome, not per month.** — Ongoing
  A monthly cap tells you when you are in trouble; tracking cost per task tells you whether it was worth it.
  Done when: your AI line item is dollars per outcome ($/report, $/ticket) — not a flat monthly figure.

If you only do one thing: **ask for cost per outcome.** Answering it forces the same tier-and-volume breakdown that routing and metering need — the rest follows from there.

The five levers

Applied in order to a $525,000 baseline, these took it to roughly $137,000 — without removing a capability from anyone.

- [ ] **Right-size the model** — −40%
  The test: of all your business tasks, how many genuinely need the best model on earth? Route the rest to a cheaper tier. The best open-weight models are within a couple of points of the paid ones on routine work.
- [ ] **Script the plain steps** — −15%
  The test: if you can write the rule down, you don't need to rent judgment to follow it. Where the task genuinely needs judgment or language, the model earns its price.
- [ ] **Engineer the context** — −25%
  The test: ask what your agent re-reads on every step. Capping it — a short rolling memory instead of the whole history — is a setting, not a rebuild. Ask your vendor about prompt caching in the same conversation.
- [ ] **Stop redoing work** — −15%
  The test: where do finished prompts and outputs live? A prompt someone tuned for four hours is an asset, and so is the artifact it produced. Both belong in shared, versioned storage.
- [ ] **Run commodity work locally** — −20%
  The test: which workloads are routine, high-volume, and sensitive? Those are the candidates. Two or three small specialist models share one GPU, and nothing leaves the building.

Check the volume before you buy hardware. Local wins on volume, not by default — against commodity-tier cloud pricing the electricity alone can cost more than the API call.

And the one that does need a budget: training. 63% of women report a lack of skills and access to AI training at work, and the C-suite expects 46% of workers to need reskilling within three years. Every lever above needs somebody who knows to pull it.

The numbers worth keeping

Everything below assumes GPT-5 standard-tier pricing — **$1.25 per million input tokens, $10 per million output** — with a summary at 10% of input length. Prices move; the ratios hold. Re-check the two rates before you quote anything.

What things weigh

ARTIFACT	TOKENS
1,000 tokens	**≈ 750 words**
…which is also	**≈ 4,000 chars**
1 PDF page	**650**
1 minute of speech	**850**
1,000 words	**1,330**
1-hour meeting	**10,000**
100-page report	**70,000**

Images are priced by area, not content: roughly (width × height) ÷ 750.

Pro tip: a dense block of text, rendered as a compact image instead of raw text, can cut input tokens by up to 3x.

What things cost

IF YOU HAVE…	BUDGET ABOUT…
A single email	**< $0.001**
A 10-page report	**≈ 2¢**
A one-hour meeting	**≈ 2¢**
A 100-page report	**≈ 16¢**
A day of email (50)	**≈ 5¢**
1,000 × 100-page reports	**≈ $158**

Read it in, hand a summary back — that is the whole bill for one item.

What scale does

Same 600 people. Same 250 working days. Same price list. The only variable is how the tool is used.

PATTERN	INPUT / PERSON / DAY	OUTPUT	PER YEAR	PER PERSON
Assistant — chat, summarize, draft	20,000	5,000	**$11,250**	**$18.75**
Agentic — it works on your behalf	500,000	20,000	**$123,750**	**$206.25**
Always-on — agents that never stop watching	2,000,000	100,000	**$525,000**	**$875.00**

Agents read far more than they write. Nothing about the price list changes between these rows — the tool simply starts reading on your behalf. That is the entire cost story.

Never goes in the prompt

A short list, worth circulating to the team once. It costs nothing to strip these out — and a great deal to explain why they were left in the prompt.

- ☐ **API keys, passwords, tokens**
- ☐ **Customer names, emails, card data**
- ☐ **Client contracts and their IP**
- ☐ **Unreleased financials**

Pair it with a shared, sanitized prompt template for the two or three workflows everyone actually runs.

Five questions to ask before you sign

? **What is the cost per outcome, not per seat or per month?** If it can't be quoted or calculated on the spot, it isn't being tracked.

? **Can I see usage by team and by workflow?** Metering should not be an upgrade.

? **Which model tier runs each task, and can I change it?** If everything runs on the frontier model, you are paying a tax on habit.

? **What do you cache, and what does the discount look like?** Repeated context is usually the largest single line in an agent bill.

? **What sits in the path of every model call, and what does it retain?** Gateways and proxies see every prompt and every response.

The one-sentence version. Tokens are cheap and volume drives the bill — and roughly three-quarters of that bill is decision-driven, not a set price.

Nick McCarty · Upskilled Consulting
nick@upskilled.consulting · linkedin.com/in/nicholasmccarty

Figures for the five levers are a working model, not a measurement — your mix will differ.

Slides, this checklist and how to reach me