

AI COST AND QUALITY AUDIT

Where Example SaaS Co.'s AI money goes, and what to change

Data reviewed: billing and usage exports, June to August 2026, a list of 11 AI tools and plans, and 25 redacted prompts.

Prepared for
Example SaaS Co.
(fictional)

Delivered
Day 5 after data

Length
11 pages + sheet

Illustrative sample. Example SaaS Co. is a fictional company and every figure here is invented to show the format. Provis has not yet completed client audits, so none of this is a result.

CONTENTS

What is in this report

1	Cover and scope	1
2	Contents	2
3	Executive summary	3
4	Where the money goes	4
5	Findings register	5
6	Finding F1 in full	6
7	Prompt recommendations	7
8	Model routing and caching plan	8
9	Keep, and what not to do	9
10	Your data, and the 30-day plan	10
11	Method, limits and confidence	11

Each finding comes with a **saving range**, the **effort** to make the change, the **risk to output quality**, how **confident** we are, and **who** can make the change. The calculation sheet behind every figure comes with the report.

EXECUTIVE SUMMARY

One page for your CFO

MONTHLY AI SPEND

\$18,400

API usage plus 11 AI tools and plans

SAVING RANGE

\$4,900–6,300

a month, if all six findings are applied. A sample, not a forecast

QUALITY RISK

Low

Two findings low, four none; each tested before switching

The short version. Most of the spend sits in two features: support ticket tagging and the in-app assistant. Both run on a larger model than they need for part of their work, and both resend the same instructions on every call.

Top three actions

ACTION	SAVING / MONTH	EFFORT
Move ticket tagging to a smaller model (F1)	\$2,100–2,600	Low
Cache the assistant's system prompt (F2)	\$900–1,200	Low
Send fewer retrieved chunks to search answers (F3)	\$700–1,000	Medium

Saving by finding (midpoint of each range, per month)

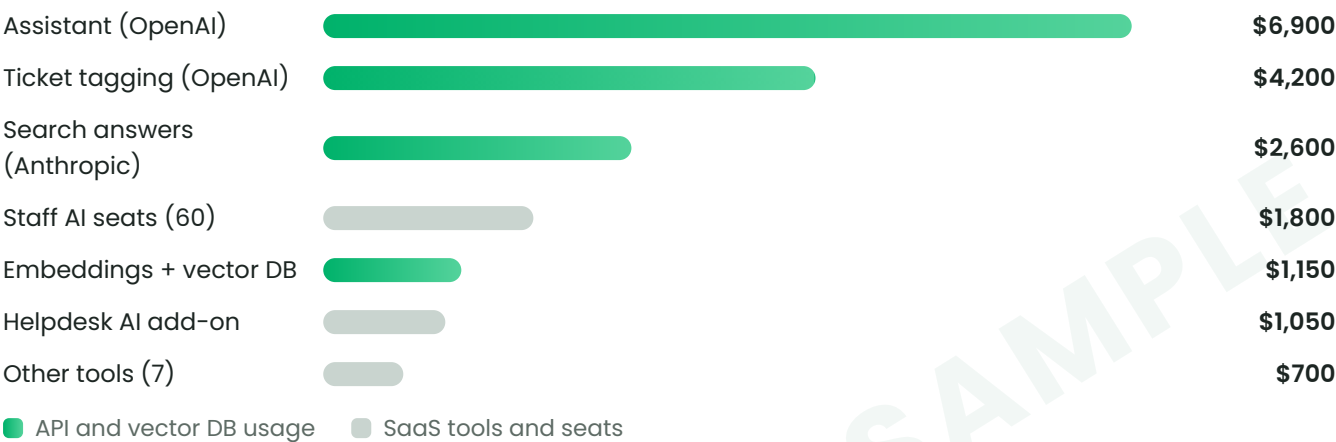
F1 Tagging model		\$2,350
F2 Prompt caching		\$1,050
F3 Fewer chunks		\$850
F4 Duplicate tool		\$475
F6 Retries		\$475
F5 Unused seats		\$415

Keep: search answers earn their cost; F3 only trims the chunks they send. **Don't:** switch helpdesk vendor to save money; see page 9.

SPEND BREAKDOWN

Where the money goes

Monthly average, June to August 2026, by feature and tool. API costs are volume times each provider's published price on the audit date.



BY PROVIDER

OpenAI 60% · Anthropic 14% · Vector DB 6% · SaaS tools and seats 19%. Rounded to the nearest 1%.

BY ENVIRONMENT

Production 88% · Staging 9% · Developer keys 3%

Staging uses the same large model as production. That is **\$1,650 a month** spent testing at production prices; see F1.

FINDINGS REGISTER

Six findings, ranked by saving

#	FINDING	SAVING / MONTH	EFFORT	QUALITY RISK	WHO
F1	Ticket tagging runs on a large model; a smaller one matches it	\$2,100 to \$2,600	Low	Low	Your team
F2	The same 1,900-token system prompt is sent on every chat call	\$900 to \$1,200	Low	None	Your team
F3	Search answers send 12 retrieved chunks; 5 cover 97% of answers	\$700 to \$1,000	Medium	Low	Us
F4	Two tools summarise the same helpdesk tickets	\$450 to \$500	Low	None	Vendor
F5	14 of 60 staff AI seats unused for 60+ days	\$400 to \$430	Low	None	Your team
F6	Timeouts retry the whole chain, doubling some calls	\$350 to \$600	Medium	None	Us

Confidence for each finding is on page 11. Savings are ranges because they depend on volume, which changes month to month.

FINDING F1

Ticket tagging runs on a large model; a smaller one matches it

EVIDENCE

41,000 tickets a month, average 1,350 input and 40 output tokens each, on the large model in production and staging.

THE CHANGE

Route tagging to the smaller model in the same family. Keep the large model as a fallback when confidence is low.

SAVING

\$2,100–2,600 / month

Range covers 35,000 to 47,000 tickets a month.

QUALITY CHECK

200 past tickets re-run on the smaller model: **196 of 200** got the same tag. The 4 that differed were ambiguous for a human too.

EFFORT	QUALITY RISK	CONFIDENCE	WHO	ROLL BACK
Low 1 day	Low	High	Your team	One config value

Monthly cost of tagging

Today (large model)		\$4,200
After (small, with fallback)		\$1,600–2,100

How to test it yourself: run 2 weeks with 10% of tickets on the smaller model and compare tag agreement before moving all traffic.

PROMPT RECOMMENDATIONS

Prompt changes, costliest first

Assistant system prompt (F2). 1,900 tokens sent on every call, including 600 tokens of examples the model no longer needs.

- You are a helpful assistant for Example SaaS Co. Always be polite, friendly and helpful. Here are 12 examples of good answers: ...[600 tokens]...

- Remember to always be polite.

+ You answer questions about Example SaaS Co. using the context provided. Be brief and polite.
+ [3 examples, 140 tokens] + cache this block

Search answers (F3). Retrieve 12 chunks, then send the top 5 by relevance score.

+ top_k: 12 → rerank → send 5

Ticket tagging (F1). Return the tag only, not a sentence of explanation: 40 output tokens become 6.

Every change is tested on your own past inputs before we recommend it; see page 11.

ROUTING AND CACHING

Which task goes to which model

TASK	TODAY	RECOMMENDED	WHY
Ticket tagging	Large model	Small model, large as fallback	196/200 same tags
Assistant answers	Large model	Keep large model	Quality dropped on the small one
Search answers	Mid model	Keep, send fewer chunks	Context is the cost, not the model
Staging tests	Large model	Small model or a local open model	No users see staging

Caching. Cache the assistant's system prompt and the search instructions. About 70% of input tokens on those calls repeat word for word.

Batching. Re-embed the help centre nightly in one batch instead of on every edit.

Open models. At this volume, hosting your own model costs more than it saves. Revisit above roughly 4 times today's tagging volume.

KEEP, AND WHAT NOT TO DO

What earns its cost

KEEP	WHY
Search answers	Users who search with it find answers faster. Keep the feature; F3 only trims the chunks it sends.
Assistant on the large model	Answer quality dropped on the smaller model in the sample test.
Helpdesk AI add-on	Drop the duplicate summariser instead (F4).

What we recommend you don't do

Don't switch helpdesk vendor to save on AI. The AI saving is about \$450 a month; migration and retraining would take more than 14 months to pay back.

Don't fine-tune a model for tagging. The smaller model already matches; fine-tuning adds cost and upkeep for no gain here.

Not covered by this audit

Code review or security testing	Not part of the audit; we did not see your code.
Production systems	No access was given or needed.
Vendor contract negotiation	We note renewal dates; you negotiate.

If your spend is already sensible, your report will say so.

YOUR DATA

What we received, and when it is deleted

RECEIVED	USED FOR
Billing and usage exports, June–August	Spend breakdown, findings
List of 11 AI tools and plans	Seats and duplicate tools
25 prompts, customer data removed	Prompt recommendations

Not received, not needed: API keys, passwords, source code, production or admin access, customer data, PHI. Files are deleted within 30 days of this report, sooner on request.

30-DAY PLAN

WEEK	DO	WHO
1	Cache prompts (F2); remove unused seats (F5)	Your team
2	10% of tagging on the small model (F1)	Your team
3	Fewer chunks for search (F3); fix retries (F6)	Us
4	All tagging moved if agreement holds; drop duplicate summariser (F4)	Your team, vendor

METHOD, LIMITS AND CONFIDENCE

How the numbers were worked out

Cost. Token volume from your usage exports times each provider's published price on the audit date. Tool costs from your invoices. The calculation sheet shows every formula.

Quality. Before recommending a cheaper option, we re-run a sample of your own past inputs on it and compare the results. Where quality drops, we say so and recommend keeping the current setup.

FINDING	CONFIDENCE	BASED ON
F1	High	200-ticket re-run
F2	High	Token counts
F3	Medium	50-question re-run
F4, F5	High	Invoices, seat logs
F6	Medium	Error logs, one month

Limits. Three months of data; seasonal peaks may differ. Savings assume today's prices and volume. We did not review code or production systems.

Want a report like this for your own AI spend? Request the free AI cost audit at provistechnologies.com/free-audit/ai-cost. Five working days after your data arrives.