BEZALEEL
CONSULTING GROUP
INTELLIGENT SYSTEMS • BUILT TO SCALE

CLAUDE ENTERPRISE TOKEN MANAGEMENT PLAYBOOK

Six levers to stop runaway AI spend. Built for the people who pay the bill.

Enterprise Claude is powerful, but cost compounds when teams don't manage tokens. These six controls — combined — typically cut spend 40–70% without sacrificing quality.

90%	17×	85%	3–5×
savings on cached content	cost gap: large vs. small models	of queries fit smaller models	budget overrun without limits

1 TURN ON PROMPT CACHING

The single highest-ROI move most teams haven't made

Claude reprocesses the same system prompts and reference docs on every call unless you cache them. Place stable content at the start of your prompt and add a cache breakpoint on the last unchanging block. **~90% savings on cached input**

- Cache system instructions, role definitions, and large reference documents
- Place cached blocks at the prompt's beginning — order matters
- Set the breakpoint on the last block that stays identical across requests

2 ROUTE BY TASK, NOT BY HABIT

Stop sending every query to your most expensive model

Haiku costs roughly a third of Sonnet and runs faster. Reserve premium models for analysis and synthesis; send classification, extraction, and triage to Haiku. About 85% of enterprise queries fit the smaller tier with no measurable quality loss. **40–60% cost reduction**

- **Haiku:** classification, tagging, extraction, simple Q&A, routing decisions
- **Sonnet:** drafting, analysis, multi-step reasoning, most enterprise workflows
- **Opus:** strategic synthesis, complex research, executive-grade output only

3 CONTROL CONTEXT BLOAT

Conversations don't shrink — they compound

Every turn re-sends the full history. By turn ten, a chat reprocesses 4,500+ tokens before reading the new question. Summarize past 8–10 turns; start fresh sessions for new topics.

- Use **/compact** after each work phase
- Use **/clear** between unrelated tasks
- Build CLAUDE.md files for stable project context

4 SET BUDGETS & ALERTS

The control most companies skip — and regret

Without per-user and per-team caps, a 250-person org routinely spends 3–5× its intended budget by month two. Configure thresholds before, not after.

- Per-user token limits aligned to role
- Per-team monthly budgets at 80% of provider cap
- Automated alerts at **50% / 80% / 100%**

5 MAKE SPEND VISIBLE

You can't manage what nobody sees

Publish a weekly dashboard of token spend by team, use case, and model. When people see the bill, behavior changes within one billing cycle.

- Track input vs. output tokens — output costs 3–5× more
- Flag outlier workflows burning tokens unnecessarily
- Enable chargeback to business units

6 TRAIN THE HUMANS

Highest-leverage investment most teams skip

Only 28% of employees feel adequately trained on their AI tools. A two-hour internal workshop moves the needle more than any new platform purchase.

- Teach what a token is — and why output costs more
- Show how conversation history compounds cost
- Publish a short "Claude at our company" prompt guide

⚡ DAILY CHECKLIST FOR CLAUDE ENTERPRISE ADMINS

- ✓ Review yesterday's top 5 token-spend workflows
- ✓ Audit model selection — anything on Opus that doesn't need it?
- ✓ Investigate any user above 2× their team's median spend
- ✓ Confirm caching is active on production system prompts
- ✓ Check budget burn rate vs. month-to-date target
- ✓ Refresh the team dashboard and post the weekly summary

> *"Token waste isn't a budget problem dressed up as a tech problem. It's a knowledge problem dressed up as a budget problem — and knowledge problems can be fixed in a quarter.*

AUTOMATE • OPTIMIZE • SCALE

Partner with us to turn your Claude deployment into a measurable competitive advantage.

CONNECT WITH US

✉ contact@bezaleelconsulting.com
🌐 bezaleelconsulting.com
✆ 800-570-6185 • in linkedin.com/in/bezaleelgroup