Google Cloud

Webinar | AMER & EMEA

Community TechTalks

November 5, 2026

AI Tokenomics in Action

Enforcing Tiered Token Budgets with Apigee

https://goo.gle/Recordings

Apigee

Housekeeping Notes

- 🎥 This session is being recorded so you can revisit it on demand later.
- 🔇 Please remain muted during the presentation
- 💬 Feel free to share questions anytime through the chat.
- 🎤 We'll open the floor for discussion and live questions **at the end** of the session.
- 👍 Use reactions and emojis to engage throughout the presentation.
- 🤝 Please help us keep the space respectful and welcoming for everyone.

Speaker



Tarun Abraham

Customer Engineer, Apigee, Google Cloud



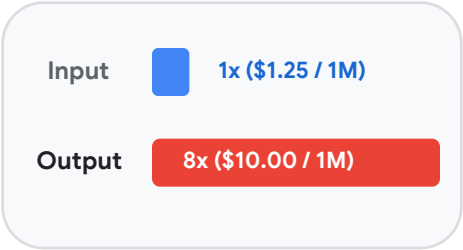
AI Tokenomics in Action

Enforcing Tiered Token Budgets with Apigee

Why LLM Inference Spend Spirals

01

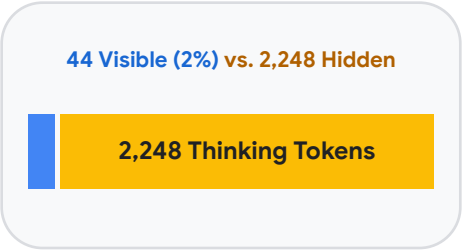
Asymmetric I/O Pricing



Output tokens cost 3x–8x more than input.

02

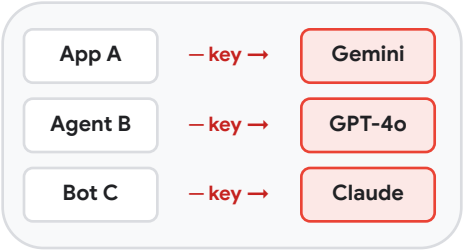
Hidden Thinking Tokens



Reasoning tokens are billed as output.

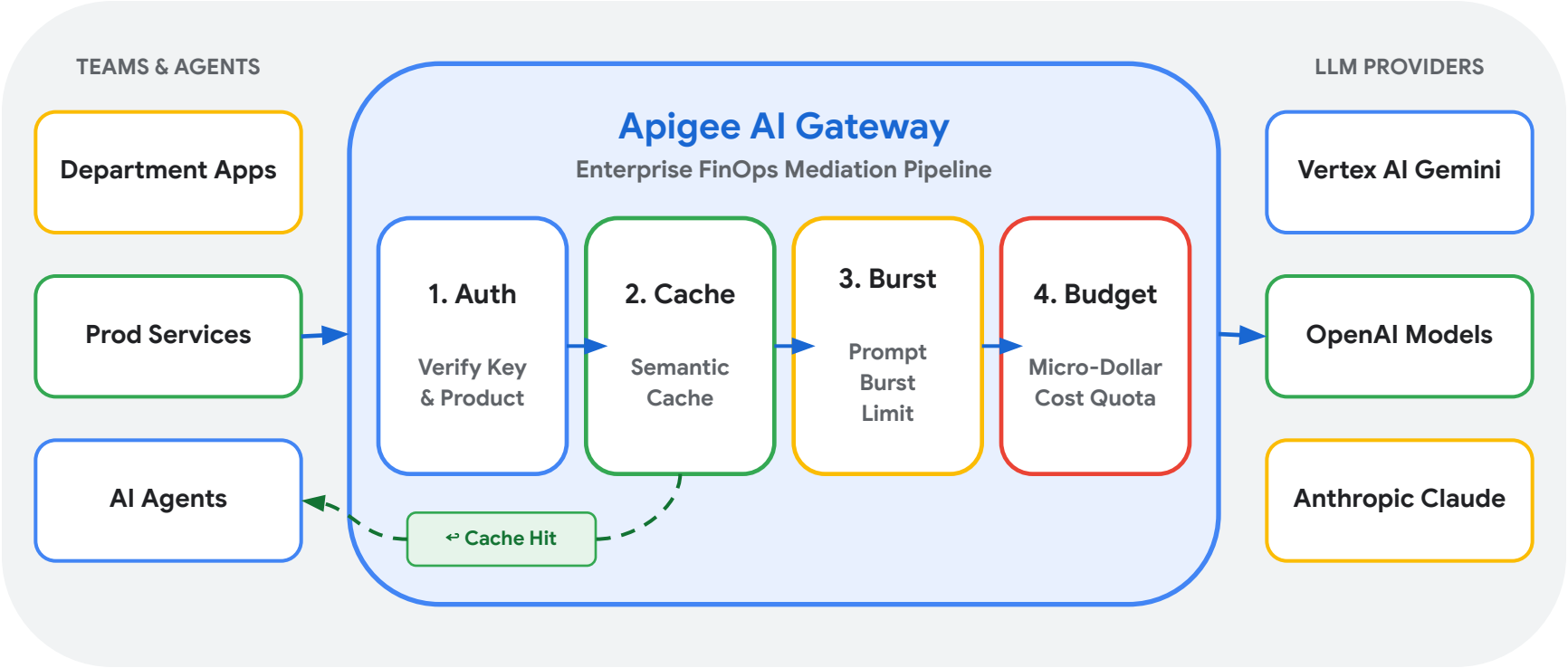
03

Multi-Team Sprawl

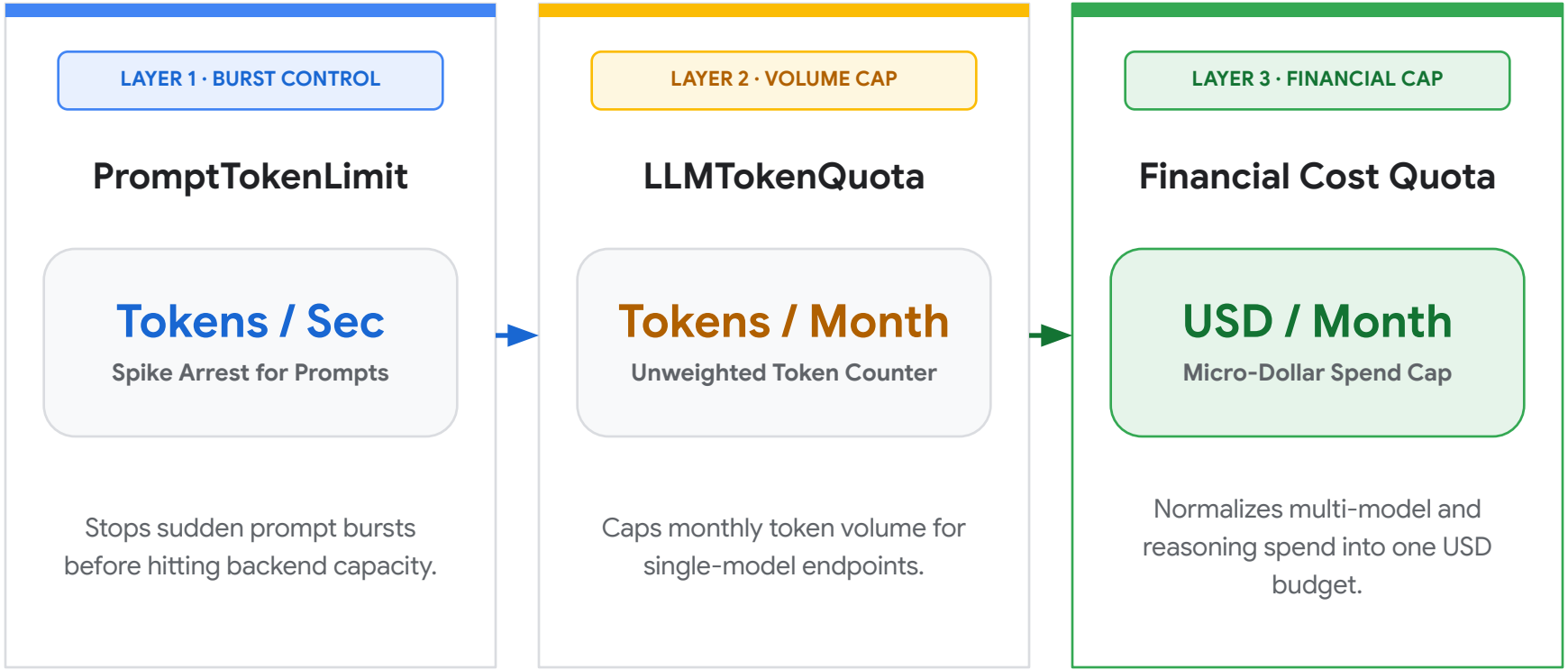


Direct provider keys bypass central budgets.

Apigee as the Central Enterprise AI Mediation Plane

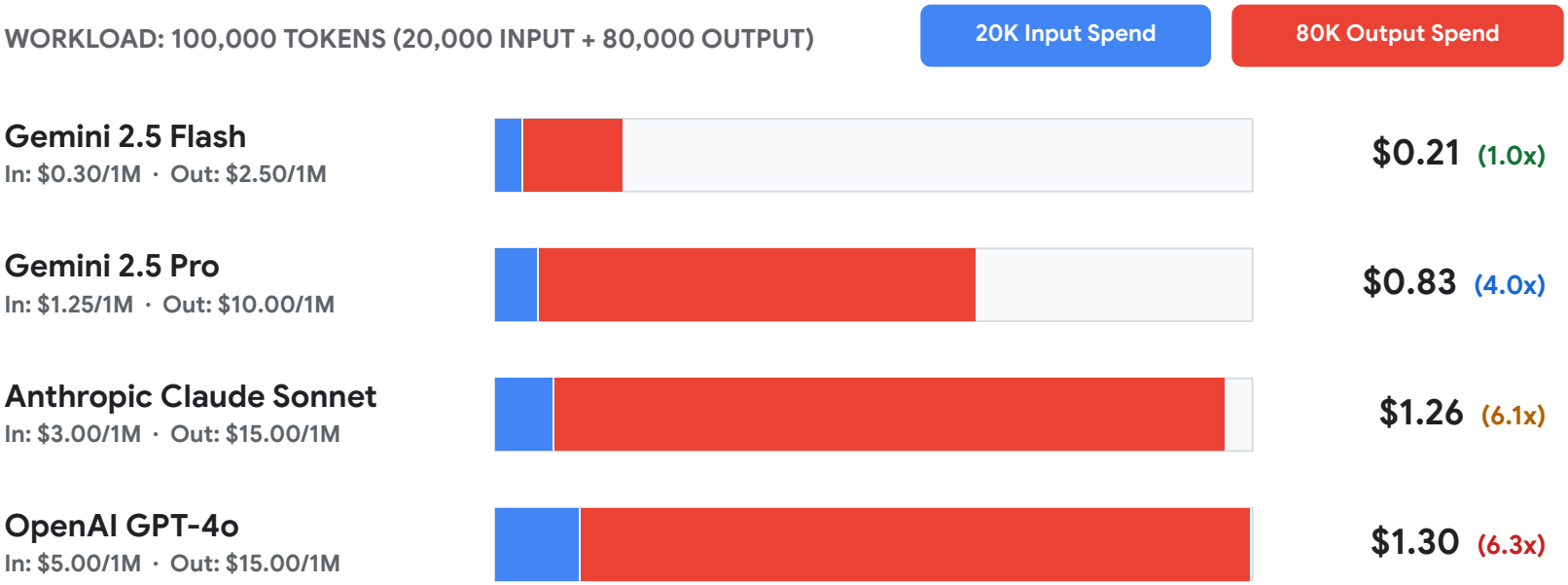


Token Limits vs. Cost Quotas: 3-Layer Defense



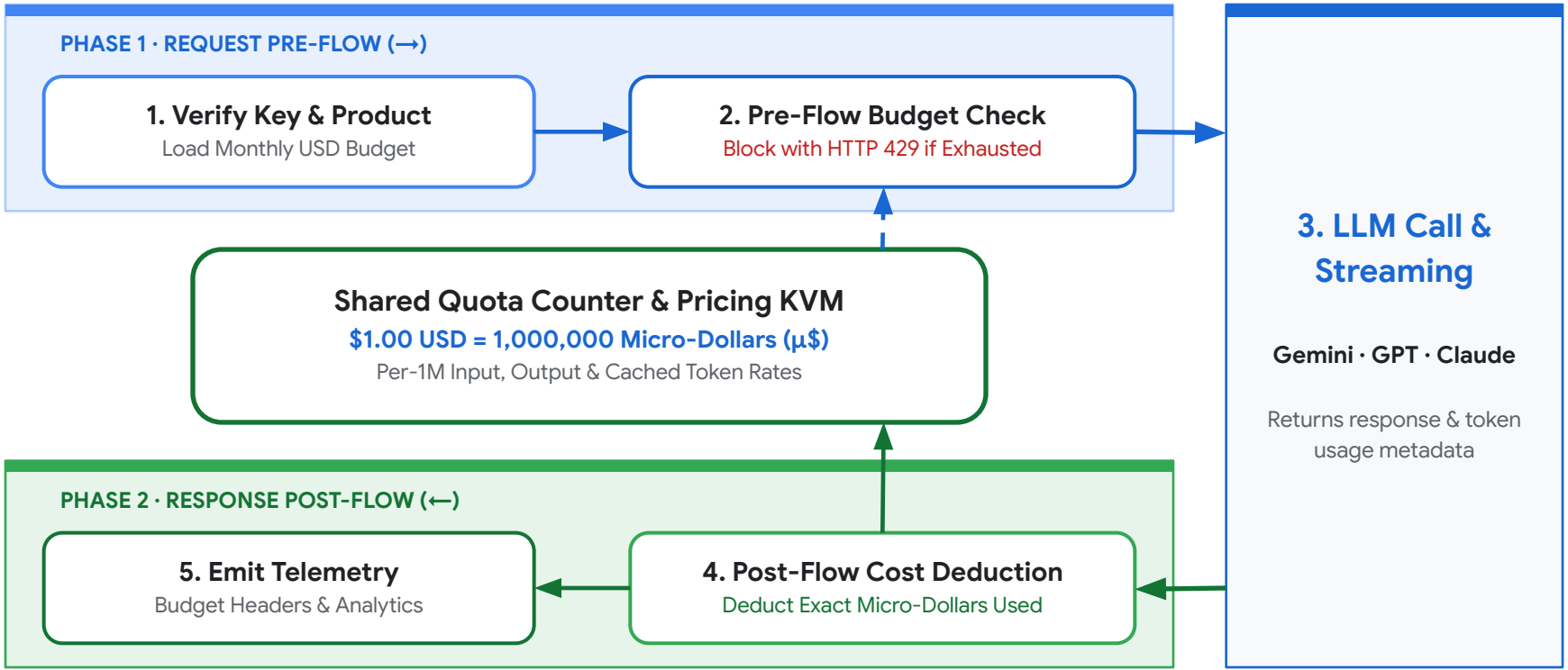
Same 100,000 Tokens, Over 6x Difference in Actual Spend

WORKLOAD: 100,000 TOKENS (20,000 INPUT + 80,000 OUTPUT)

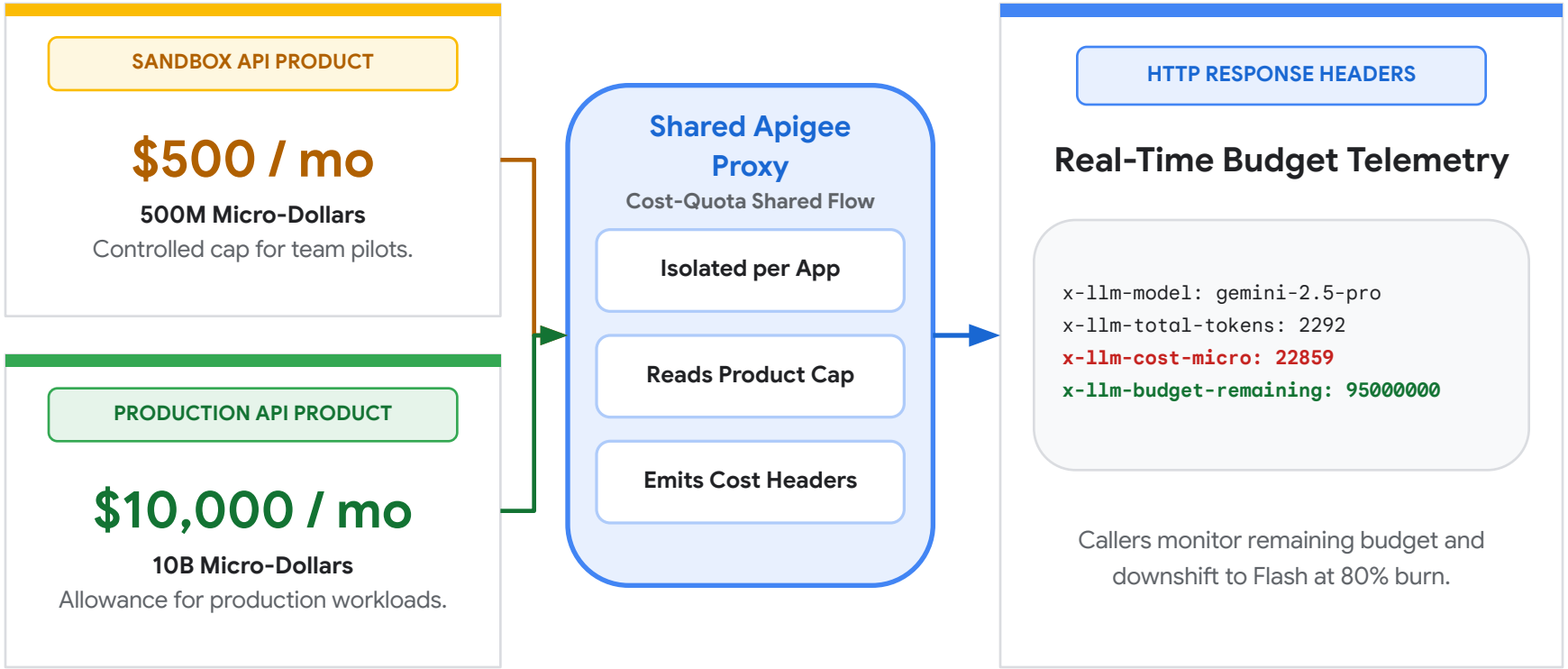


Illustrative published per-1M token rates

Reference Architecture: Reusable Cost-Quota Shared Flow



Multi-Tenant Governance with Apigee API Products



Runtime Budget Lifecycle in Action

SANDBOX TIER PROGRESSION · $500.00 MONTHLY CAP

STAGE 1 · IN-BUDGET

200 OK

$499.98 / $500

Post-Flow deducts exact spend.

STAGE 2 · CAP CROSSED

200 OK

$500.01 / $500

In-flight call finishes cleanly.

STAGE 3 · EDGE CUTOFF

HTTP 429

$500.01 Flatlined

Next call blocked at edge.

PARALLEL TENANT · ISOLATED LANE

Production Tier (200 OK)

$6,400 / $10,000 Used

Isolated counter continues unaffected.

FinOps Observability & Department Chargeback

Chargeback by Team

Attribute USD spend per API Product.

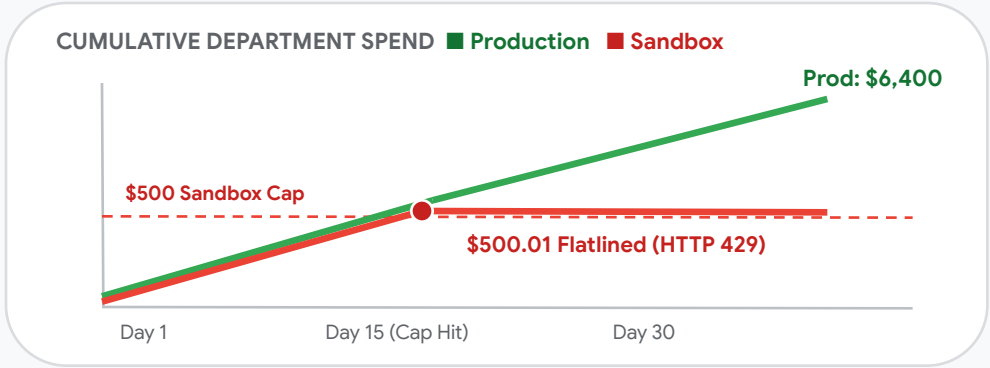
Spend by Provider

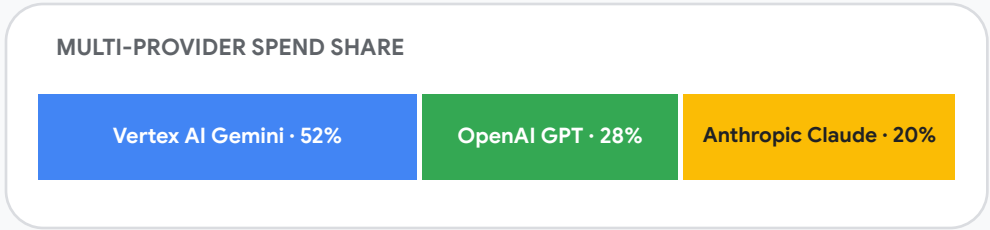
Compare Gemini, OpenAI, and Claude.

Quota Cutoff Proof

Verify spend flatlines at the cap.

CUSTOM ANALYTICS REPORT · CUMULATIVE SPEND & PROVIDER SHARE





Production Rollout Checklist & Get Started

01

Product Tiers

Set monthly USD caps on API Products.

02

Encrypted KVM

Store per-million model token rates.

03

Shared Flow

Enforce Pre-Flow checks and Post-Flow charges.

04

Smart Routing

Downshift to Flash at 80% budget burn.

REFERENCE IMPLEMENTATION & STARTER KIT

github.com/tarunabraham1986/apigee-llm-budget-control

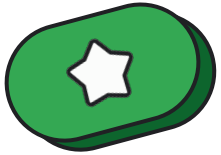
Q&A

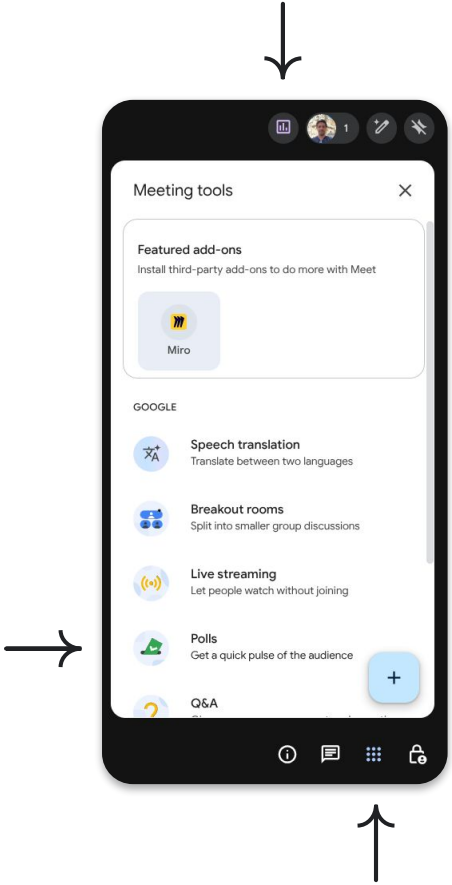


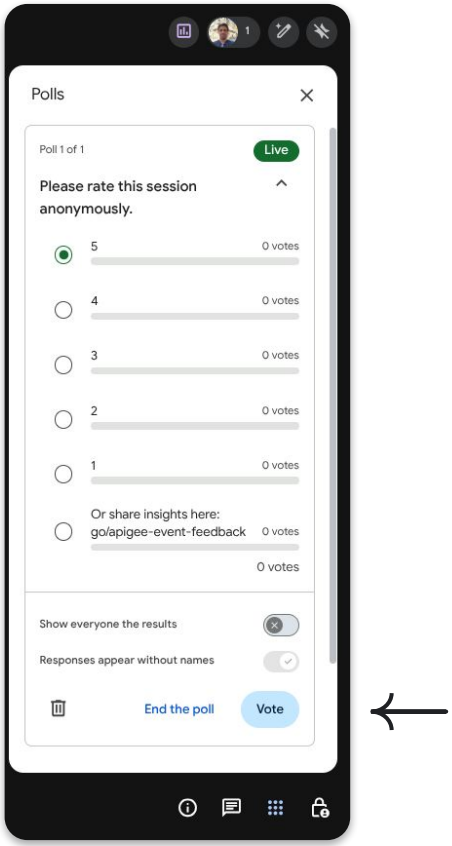
Please share your feedback!

Help us make our sessions even better!

Please provide your **anonymous feedback** by rating this session.







Catch up!

New Community Content

- Building an adaptive AI gateway with Apigee for stateful multi-agent workloads ↗
- Apigee Hybrid 1.17.0 introduces Model Context Protocol (MCP) support ↗

Upcoming Events

- Nov 12 | Leveraging "Antigravity" to Architect High Performance Semantic Model Routers in Apigee
- Nov 19 | Security for AI workloads with Model Armor
- Dec 10 | Gouvernance de l'IA en temps réel, FinOps et routage avec Apigee

Google Cloud

Community



Thank You