TYGART MEDIA · CLAUDE & ANTHROPIC REFERENCE

How to Run Claude Code on Vertex AI Using Your GCP Credits

Edition of November 13, 2026 — facts verified November 13, 2026. Current edition always at https://tygartmedia.com/claude-code-vertex-ai-gcp/.

Direct Answer (November 13, 2026): Set `CLAUDE_CODE_USE_VERTEX=1`, `CLOUD_ML_REGION=global`, and `ANTHROPIC_VERTEX_PROJECT_ID` (or run `/setup-vertex` inside Claude Code) to route Claude Code billing through Google Cloud Vertex AI using GCP credits instead of an Anthropic API key (Claude Code docs). Enable the current **Sonnet 5.5 / Opus 5.5 / Haiku 5.5** models available in your Vertex Model Garden catalog.

Last refreshed: November 13, 2026 · Sources: Anthropic models overview & Claude Code docs

Model Accuracy Note — Updated November 13, 2026

Current Anthropic lineup (models overview): **Fable 5.2** (`claude-fable-5-2`, $8/$40) · **Opus 5.5** (`claude-opus-5-5`, $4/$20 daily driver) · **Sonnet 5.5** (`claude-sonnet-5-5`, $2/$10) · **Haiku 5.5** (`claude-haiku-5-5`, $0.10/$0.50 ≤100K) — all 1M context / 128K output. Opus 4.7, Sonnet 4.6, and Haiku 4.5 in the Vertex pinning snippets below reflect April–May 2026 Vertex IDs; check Vertex Model Garden for the latest 5.5/5.2 model IDs enabled in your project. See current model tracker →

Claude AI · Tygart Media

Claude Code is Anthropic's terminal-native coding agent. By default it bills through your Anthropic account. But you can route it entirely through Google Cloud's Vertex AI — meaning it charges your GCP account instead, and you can use existing GCP credits, startup credits, or free trial credits to run it at no incremental cost. Here's the exact setup.

What You Need Before Starting

What you need before starting.

A Google Cloud account with a project created. Vertex AI API enabled on that project. Claude models requested and approved in Vertex AI Model Garden. Claude Code installed (`npm install -g @anthropic-ai/claude-code`). The `gcloud` CLI installed and authenticated. That's it — no Anthropic API key required once this is configured.

Need this set up for your team?

I set up Claude integrations, GCP infrastructure, and AI workflows for businesses. If you'd rather ship than configure — will@tygartmedia.com

Step 1: Enable Vertex AI and Request Claude Model Access

Step 1 — enable Vertex AI and request Claude model access.

In the Google Cloud Console, go to **Vertex AI > Model Garden** and search for "Claude." Request access to your primary coding model (current Anthropic recommendation: **Claude Sonnet 5.5** and **Claude Opus 5.5**, or legacy Sonnet 4.6 where pinned) plus a fast lightweight tier (**Claude Haiku 5.5**, or Haiku 4.5 if 5.5 is not yet enabled in your region) in Vertex AI Model Garden.

Model access approval is typically instant for most GCP accounts.

Step 2: Authenticate with Google Cloud

Run both commands below — the first authenticates your user account, the second sets application default credentials that Claude Code will pick up automatically:

gcloud auth login

gcloud auth application-default login

Set your project: `gcloud config set project YOUR-PROJECT-ID`

Enable the Vertex AI API: `gcloud services enable aiplatform.googleapis.com`

Step 3: Configure Claude Code to Use Vertex AI

Step 3 — configure Claude Code to use Vertex AI.

Set these environment variables. On macOS/Linux, add them to your `~/.zshrc` or `~/.bashrc`. On Windows, use PowerShell's `[System.Environment]::SetEnvironmentVariable` at the User level so they persist across sessions.

macOS / Linux:

export CLAUDE_CODE_USE_VERTEX=1

export CLOUD_ML_REGION=global

export ANTHROPIC_VERTEX_PROJECT_ID=your-project-id

export ANTHROPIC_DEFAULT_SONNET_MODEL=<your-verified-vertex-sonnet-id> # e.g., claude-sonnet-4-6 or newer Model Garden Sonnet 5.5 ID

export ANTHROPIC_DEFAULT_HAIKU_MODEL=<your-verified-vertex-haiku-id> # e.g., claude-haiku-4-5@20251001 or newer Model Garden Haiku 5.5 ID

Windows (PowerShell — run once, persists across sessions):

[System.Environment]::SetEnvironmentVariable("CLAUDE_CODE_USE_VERTEX","1","User")

[System.Environment]::SetEnvironmentVariable("CLOUD_ML_REGION","global","User")

[System.Environment]::SetEnvironmentVariable("ANTHROPIC_VERTEX_PROJECT_ID","your-project-id","User")

Step 4: Verify the Setup

Launch Claude Code and run `/status`. You should see `API provider: Google Vertex AI` and your GCP project ID. If you see the Anthropic API provider instead, your environment variables haven't loaded — restart your terminal and try again.

Step 5: Use the New Wizard (Claude Code v2.1.98+)

If you're on Claude Code version 2.1.98 or later, you can skip manual environment variable setup. Run `/setup-vertex` inside Claude Code and the wizard walks you through project selection, region, and model pinning automatically. Run `claude --version` to check your version first.

Region Selection: Global vs Regional Endpoints

Use `CLOUD_ML_REGION=global` unless you have specific compliance reasons to pin to a region. Global endpoints get the latest models first, have better availability, and don't incur the 10% regional pricing premium. If you need data residency in a specific geography, use `us-east5`, `us-central1`, or `europe-west1` — but verify your target Claude models are available in that region first, as not all models are available in all regions.

Model Pinning for Teams

If you're deploying Claude Code to multiple team members, pin specific model versions rather than using aliases. Model aliases like "sonnet" resolve to the latest version, which may not be enabled in your Vertex AI project when Anthropic ships an update. Pinning prevents silent failures on update day:

export ANTHROPIC_DEFAULT_SONNET_MODEL=claude-sonnet-4-6 # substitute current Model Garden Sonnet 5.5 ID once enabled

export ANTHROPIC_DEFAULT_HAIKU_MODEL=claude-haiku-4-5@20251001 # substitute Model Garden Haiku 5.5 ID once enabled

When to Choose Vertex AI vs. the Direct Anthropic API

Both paths run identical Claude models (1M context window and 128K output across Fable 5.2, Opus 5.5, Sonnet 5.5, and Haiku 5.5 once enabled in Model Garden). Choose between them based on cloud governance vs. setup speed:

- **Choose Vertex AI on GCP for:** consolidated GCP billing, existing GCP/startup credits, Google Cloud IAM + Cloud Audit Logs instead of standalone API key distribution, **VPC Service Controls** (keeping inference inside a private cloud perimeter rather than traversing the public internet), and strict regional data residency (EU/US pinning).
- **Choose the Direct Anthropic API for:** 5-minute key setup with no GCP project/IAM overhead and **day-one model access** (new Claude releases reach the direct API before Vertex AI Model Garden rollout).

Pricing Comparison: Direct API vs. Vertex AI Global & Regional

Model	Anthropic Direct	Vertex AI (Global)	Vertex AI (Regional)
Claude Fable 5.2 (in / out)	$8 / $40 per MTok	Matches Anthropic	+10% regional premium
Claude Opus 5.5 (in / out)	$4 / $20 per MTok	Matches Anthropic	+10% regional premium
Claude Sonnet 5.5 (in / out)	$2 / $10 per MTok	Matches Anthropic	+10% regional premium
Claude Haiku 5.5 (≤100K in / out)	$0.10 / $0.50 per MTok	Matches Anthropic	+10% regional premium

Common Error: 429 Resource Exhausted

If you see 429 errors after setup, your project's Vertex AI quota for Claude models needs to be increased. Go to **Cloud Console > IAM & Admin > Quotas**, filter by "anthropic," and request an increase for the models you're using. Approvals are typically fast for standard business accounts.

Can I run Claude Code on Vertex AI for free?

Yes if you have unused GCP credits. New Google Cloud accounts receive $300 in free credits. All GCP credits — startup programs, free trial, committed use discounts — apply to Claude usage through Vertex AI.

Do I need an Anthropic API key to use Claude Code on Vertex AI?

No. When configured for Vertex AI, Claude Code authenticates through your Google Cloud credentials (gcloud). No Anthropic API key is needed or used.

Is Claude Code on Vertex AI slower than the direct Anthropic API?

In practice, latency is comparable. The global endpoint routes dynamically and generally performs well. Regional endpoints may add slight latency depending on your geographic distance from the selected region.

Work with Tygart Media

Need Claude running on your GCP infrastructure?

We've deployed Claude on Vertex AI for content operations, internal tools, and automated publishing pipelines. If you're an engineering team looking to integrate Claude into your GCP environment, we can help scope and build it.

See How We Work →

Primary & Reference Links Cited on This Page

- Anthropic models overview — `https://platform.claude.com/docs/en/about-claude/models/overview`
- Claude Code docs — `https://code.claude.com/docs/en/overview`
- Tygart Media: See current model tracker →
- Tygart Media: See How We Work →