GHOST LOGICS LLC

The Four Assistants
Claude · ChatGPT · Gemini · Grok

Cost, Usage Management & Anti-Hallucination Field Guide

What each assistant actually costs, how the free/paid/API layers differ, how to manage chats efficiently, and premade guardrail prompts to keep every one of them honest.

Pricing verified August 2026. Every provider changes pricing often — treat the numbers here as a snapshot, not a permanent fact.

Table of Contents

SECTION 01

How to Read Any AI Provider's Pricing

The universal structure behind all four

Every major AI company sells access to the same underlying models through the same three layers. Once you see this pattern, understanding a new provider's pricing page takes minutes instead of feeling like starting from zero each time.

Layer 1: The Consumer Chat App

A subscription (Free, mid-tier, top-tier) that gets you a chat window with usage caps measured loosely in “messages per day/week” rather than tokens. This is what Claude.ai, ChatGPT, the Gemini app, and Grok's app/X integration all are.

Layer 2: The Developer API

Pay-per-token, no subscription, billed precisely by usage. This is what Continue connects to when you select Claude/GPT/Gemini/Grok in the model dropdown — completely separate billing from the consumer app subscriptions, even if you use the same account.

Layer 3: Business/Team Tiers

Per-seat pricing for organizations, with admin controls, higher limits, and usually a data-training opt-out guarantee that individual free/low tiers don't always include.

The Single Most Important Habit

A subscription to the consumer app (Claude Pro, ChatGPT Plus, etc.) does NOT give you API access, and API credits do NOT give you the consumer app's subscription features. They are billed, and often literally logged into, completely separately — this is the exact mix-up that already happened once tonight with an Anthropic/OpenAI billing page.

SECTION 02

Claude & Anthropic

Verified against Anthropic's own pricing/support pages, August 2026

The Three Products

- **Claude.ai** — the consumer chat app (what this conversation is happening in)
- **Claude Code** — the agentic coding tool, bundled into paid Claude.ai plans (no standalone free tier)
- **Claude API** — pay-per-token developer access, what Continue's Claude Sonnet entry uses via your own key

Claude.ai Consumer Plans

Plan	Price	What You Get
Free	$0	Sonnet/Haiku access, roughly 30-100 messages/day with a rolling reset window
Pro	$20/mo	Much higher limits, Claude Code included, Cowork, priority access to newest models
Max 5x	$100/mo	5x Pro's usage headroom — same features, more room before hitting limits
Max 20x	$200/mo	20x Pro's usage headroom — for very heavy daily use
Team Standard	~$20-25/seat/mo	Centralized billing, admin controls, no Claude Code
Team Premium	~$100-125/seat/mo	Adds Claude Code access per seat
Enterprise	Custom	SSO, larger context windows, compliance tooling

Max and Pro/Team differ mainly in usage headroom, not features — you're paying for more room before hitting a limit, not a materially different product.

Claude API (What Continue Actually Uses)

Billed per million tokens, separately from any Claude.ai subscription. As of August 2026, the current flagship (Opus 4.8) runs roughly $5 input / $25 output per million tokens; Sonnet-tier models are meaningfully cheaper. Exact current rates should always be confirmed at `console.anthropic.com` since this changes.

Where to Manage Your Account

- Billing/credits: `console.anthropic.com` → Settings → Billing
- Claude.ai subscription: `claude.ai` → Settings → Billing
- These are the same company but genuinely separate billing surfaces — API credits running low will not affect your Claude.ai chat subscription, and vice versa

SECTION 03

ChatGPT & OpenAI

Verified against current OpenAI pricing pages, August 2026

The Two Products

- **ChatGPT** — the consumer chat app
- **OpenAI API** — pay-per-token developer access, what Continue's GPT entry uses

ChatGPT Consumer Plans

Plan	Price	What You Get
Free	$0	Limited access to a lighter model, ads present in the US
Go	$8/mo	More messages/uploads than Free, still not the flagship model in standard chat
Plus	$20/mo	Flagship model access, Deep Research, image generation, Codex, Agent Mode
Pro	$100/mo	5x Plus usage headroom, more Codex/Deep Research capacity
Pro (higher)	$200/mo	20x Plus usage, largest context window, highest-effort reasoning mode
Business	~$20-25/seat/mo	Admin controls, SSO, no training on your data by default
Enterprise	Custom	Full org controls, custom deployment options

An Important OpenAI-Specific Quirk

The Model You Get Isn't Always the Model on the Tin
OpenAI's “Auto” model selector can route a routine question to a lighter/cheaper model even on a paid plan, reserving the flagship for questions it judges as harder. If you need the flagship model specifically for a given task, check whether your interface lets you select it manually rather than trusting Auto to pick it for you.

OpenAI API (What Continue Actually Uses)

Billed per million tokens. As of August 2026, the current flagship reasoning model runs roughly $5 input / $30 output per million tokens, with a mid-tier model around $2/$12 and a budget-tier model around $0.20/$1.20. Confirm exact current rates at `platform.openai.com` before relying on them for cost planning.

SECTION 04

Gemini & Google

Verified against current Google AI pricing, August 2026

The Products

- **Gemini app** — the consumer chat app
- **Google AI Studio** — a free/low-cost developer sandbox with its own separate limits
- **Gemini API / Vertex AI** — pay-per-token developer access, what Continue's Gemini entry uses

Gemini Consumer Plans

Plan	Price	What You Get
Free (AI Studio)	$0	Flash-tier models only — Pro-tier models removed from free access in April 2026
Google AI Plus	~$5-8/mo	Entry paid tier, more storage, still limited on top-tier model access
Google AI Pro	$19.99/mo	Gemini 3.x Pro access, 1M token context window
Google AI Ultra (5x)	$99.99/mo	5x Pro usage limits, Deep Think reasoning mode, large cloud storage bundle
Google AI Ultra (20x)	$199.99/mo	20x Pro usage limits, maximum headroom

Google restructured this pricing significantly at I/O 2026 — if you're seeing older screenshots or articles referencing a $249.99 Ultra tier, that's the pre-restructure price.

Gemini API (What Continue Actually Uses)

Billed per million tokens, with an unusually wide spread between the cheapest and most expensive models — a budget “Flash-Lite” tier around $0.10/$0.40 per million tokens, versus the flagship Pro tier around $2/$12 (rising further above very large context windows). This spread is wider than Claude or GPT's equivalent tiers, so model choice matters more for cost control here.

A Gemini-Specific Cost Trap

Thinking Tokens Bill as Output
Gemini's extended-reasoning (“thinking”) tokens are billed at the output rate, not a separate lower rate. A complex question that triggers heavy internal reasoning can cost more than the sticker price suggests — factor this in in when estimating Gemini API costs for reasoning-heavy tasks.

SECTION 05

Grok & xAI

Verified against current xAI pricing, August 2026

The Two Purchase Paths

Unusual among the four: Grok is sold two separate ways — directly through xAI (`grok.com`, the SuperGrok tiers) or bundled inside an X (formerly Twitter) subscription. Same underlying model access, different purchase path and price.

Grok Consumer Plans

Plan	Price	Purchased Via	Notes
Free	$0	grok.com or X	Limited prompts, no top-tier model access
X Premium	$8/mo	X	Light Grok access bundled with X features
SuperGrok Lite	$10/mo	grok.com	Entry standalone tier, region-dependent availability
SuperGrok	$30/mo	grok.com	Full current-gen model, DeepSearch, Big Brain mode, image generation
X Premium+	$40/mo	X	Same model tier as SuperGrok, bundled with ad-free X
SuperGrok Heavy	$300/mo	grok.com	Multi-agent “Heavy” mode, first/full access to newest model
Grok Business	$30/seat/mo	xAI	Team billing, no training on your data

A Genuinely Important Privacy Note

Free and Individual Paid Tiers May Train on Your Conversations
Unlike some competitors' paid tiers, xAI's free, X Premium, X Premium+, and individual SuperGrok plans may use your conversations for model training by default. There's typically an opt-out in account settings — worth checking explicitly if you ever discuss anything sensitive (proprietary code, business strategy) through Grok, rather than assuming a paid tier automatically means privacy the way it does with some other providers.

xAI API (What Continue's Grok Entry Uses)

Billed per million tokens, OpenAI-compatible format. As of August 2026, the current flagship runs roughly $2 input / $6 output per million tokens (with a steep cached-input discount), while the fast/budget-tier model runs closer to $0.20/$0.50 — genuinely one of the cheapest frontier-adjacent options available, which is part of why it's a reasonable middle ground in your routing table for quick, contained edits.

SECTION 06

Side-by-Side Cost Comparison

Consumer Chat Apps — Entry Paid Tier

Provider	Entry Paid Tier	Price
Claude	Pro	$20/mo
ChatGPT	Plus	$20/mo
Gemini	AI Pro	$19.99/mo
Grok	SuperGrok	$30/mo

API — Approximate Cheapest Usable Model (per million tokens)

Provider	Cheapest Tier	Approx. Input / Output
Gemini	Flash-Lite	$0.10 / $0.40
Grok	Fast tier	$0.20 / $0.50
ChatGPT	Budget tier	$0.20 / $1.20
Claude	Haiku tier	Generally the cheapest of Anthropic's own lineup, check current rate

API — Approximate Flagship Model (per million tokens)

Provider	Flagship	Approx. Input / Output
Claude	Opus 4.8	$5 / $25
ChatGPT	Flagship (Sol-tier)	$5 / $30
Grok	4.5	$2 / $6
Gemini	3.1/3.x Pro	$2 / $12 (rising above large context)

All figures approximate and change frequently — confirm current rates directly with each provider before relying on them for a real budget decision.

SECTION 07

Managing Chats Efficiently

Habits that apply across all four platforms

Start a Fresh Chat for a New Topic

Every message in a long conversation gets re-processed as context for the next response — a sprawling thread covering five unrelated topics costs more per message than five short, focused chats would, and often produces worse answers as genuinely relevant context gets diluted by irrelevant history. Start a new chat when the topic genuinely changes.

Use Search/Memory Features Instead of Re-Pasting Context

Claude, ChatGPT, and Gemini all offer some form of searching or referencing past conversations rather than requiring you to re-paste earlier context by hand. Using this (when available and when it's actually the right past context) is both faster and cheaper than re-explaining something you've already told the assistant in a previous session.

Use Projects/Folders for Ongoing Work

Claude Projects, ChatGPT Projects, and Gemini's equivalent grouping features let you attach standing context (a codebase description, a style guide, project-specific facts) once, rather than repeating it at the start of every new chat about that same project.

Don't Paste What You Don't Need

Pasting an entire file when only one function is relevant, or an entire log when only the last 20 lines matter, costs real tokens for no benefit — and can actually hurt answer quality by diluting the model's attention across irrelevant content. Trim before pasting when you can.

Match Verbosity to the Task

If you only need a yes/no or a one-line answer, say so explicitly. Left unconstrained, models often default to more thorough (and more expensive, on API billing) responses than the situation actually calls for.

Know When a Long Thread Should Be Closed and Restarted

Signal to Watch For

If a conversation has drifted across many unrelated subtopics, or you notice the assistant referencing something incorrectly from much earlier in the thread, that's a sign the context has gotten noisy. Summarize what you actually need carried forward, and start clean — cheaper and usually more accurate than continuing to pile onto an already-sprawling thread.

SECTION 08

Anti-Hallucination Guardrails

Premade prompts for each assistant — paste at the start of a chat

These are general-purpose guardrail preambles. Paste the relevant one as your first message (or into a Project/Custom Instructions field, where the platform supports it) before the real task, especially for factual research, technical claims, or anything where a confidently wrong answer would be costly to act on.

Universal Core (works on all four, adapt tone as needed)

```
Before answering, follow these rules:
1. If you are not confident in a factual claim, say so explicitly rather than stating it as fact.
2. Do not invent function names, API endpoints, library methods, file paths, or citations that
you have not verified exist. If you're inferring rather than certain, label it as an inference.
3. For anything time-sensitive (current prices, current events, current software versions),
flag that your knowledge may be outdated rather than presenting it as current.
4. If a question has no good answer, say so rather than producing a plausible-sounding wrong one.
5. Distinguish clearly between what you know, what you're inferring, and what you're guessing.
```

Claude-Specific Addition

Claude can search the web when the feature is enabled. Add: *“If this question involves current information, search rather than answering from memory.”* Claude is generally strong at flagging its own uncertainty unprompted, but this addition removes any ambiguity about when to search versus recall.

ChatGPT-Specific Addition

Given the Auto-routing behavior noted in Section 3, add: *“If this task benefits from your most capable reasoning mode, use it rather than a faster/lighter response.”* Also worth adding for coding tasks specifically: *“Do not fabricate package names or API signatures — if uncertain whether something exists, say so.”*

Gemini-Specific Addition

Given Gemini's very large context window, it's tempting to paste huge amounts of source material and trust it's all being weighed evenly. Add: *“If you're drawing a conclusion from only part of the material I provided, tell me which part, rather than implying you considered all of it equally.”*

Grok-Specific Addition

Grok has strong real-time search/X-data integration built in. Add: *“For anything about current events or real-time information, use live search rather than relying on training data, and tell me if you did or didn't search.”* Given the data-training note in Section 5, also worth adding when discussing anything sensitive: *“Do not treat this conversation as a source you'd reference or repeat elsewhere.”*

A Guardrail Prompt Reduces Risk — It Doesn't Eliminate It
No preamble prevents every hallucination. Treat these as reducing the rate of confident-wrong answers, not as a guarantee. For anything genuinely high-stakes (a legal claim, a medical detail, a production-breaking code change), independent verification still matters regardless of which assistant or guardrail prompt is in use.

SECTION 09

Quick Reference Cheat Sheet

Where to Check Each Balance

Provider	Consumer App Billing	API Billing
Claude	claude.ai → Settings → Billing	console.anthropic.com → Billing
ChatGPT	chatgpt.com → Settings → Billing	platform.openai.com → Billing
Gemini	Google One / AI subscription settings	aistudio.google.com or console.cloud.google.com
Grok	grok.com or X → Settings	console.x.ai → Billing

The Three Things to Remember

1. Consumer app subscriptions and API billing are always separate, even at the same company.
2. Paste the relevant anti-hallucination guardrail before anything factual, current, or high-stakes — across any of the four.
3. Every number in this manual will drift. Verify at the source before making a real budget decision, not just once and forget it.

Ghost Logics LLC ? Pricing verified August 2026 against provider-published and provider-sourced pricing

pages; confirm current rates directly with each provider before relying on this for budget decisions.