AI TOKENOMICS GUIDE

Date: August 20, 2026
Author: Justin Daniels, Baker Donelson
RE: Understanding Token-Based AI Pricing for Legal Technology

Executive Summary: What Every GC Should Know

- AI vendors such as Anthropic (Claude), OpenAI (GPT), and Google (Gemini) charge developers per "token": a small chunk of text, roughly three-quarters of a word. Legal AI platforms may purchase access on this basis behind the scenes, even when customers experience the product as a flat monthly subscription.
- Cost is driven far more by how a tool is used than by which vendor supplies it. The biggest levers are: which underlying model is used, how much text goes in, how much text (and "thinking") comes back out, and how many steps an automated task takes.
- Output (the AI's response) costs roughly four to six times more per token than input (what you send in), on every major vendor as of August 2026. Verbose answers and multistep "agentic" workflows cost more than short, simple ones, often substantially more.
- These prices are not static: at least one major vendor adjusted rates in ten of the last 13 months, making contract flexibility important.
- This guide distinguishes between figures that are directly verified vendor list prices and illustrative estimates. Final budgeting should be based on the legal AI vendor's actual pricing, discounts, and contract terms.

1. What Is a "Token": How Does This Pricing Work?

A token is a small piece of text that an AI model reads or generates – roughly three-quarters of an English word.

Pricing is split between input tokens (what you send to the model) and output tokens (what it generates).

- **Input tokens:** Your question, instructions, and any attached documents or prior context.
- **Output tokens:** The model's visible answer and any billed internal reasoning or "thinking."

Legal AI platforms are built on this layer, typically combining one or more underlying models with their own interface and workflows.

2. The Variables That Actually Drive Cost

Two prompts can look similar to a person and cost very differently to run, depending on the variables below. This table is meant to understand the variables and their implications as opposed to a precise pricing framework.

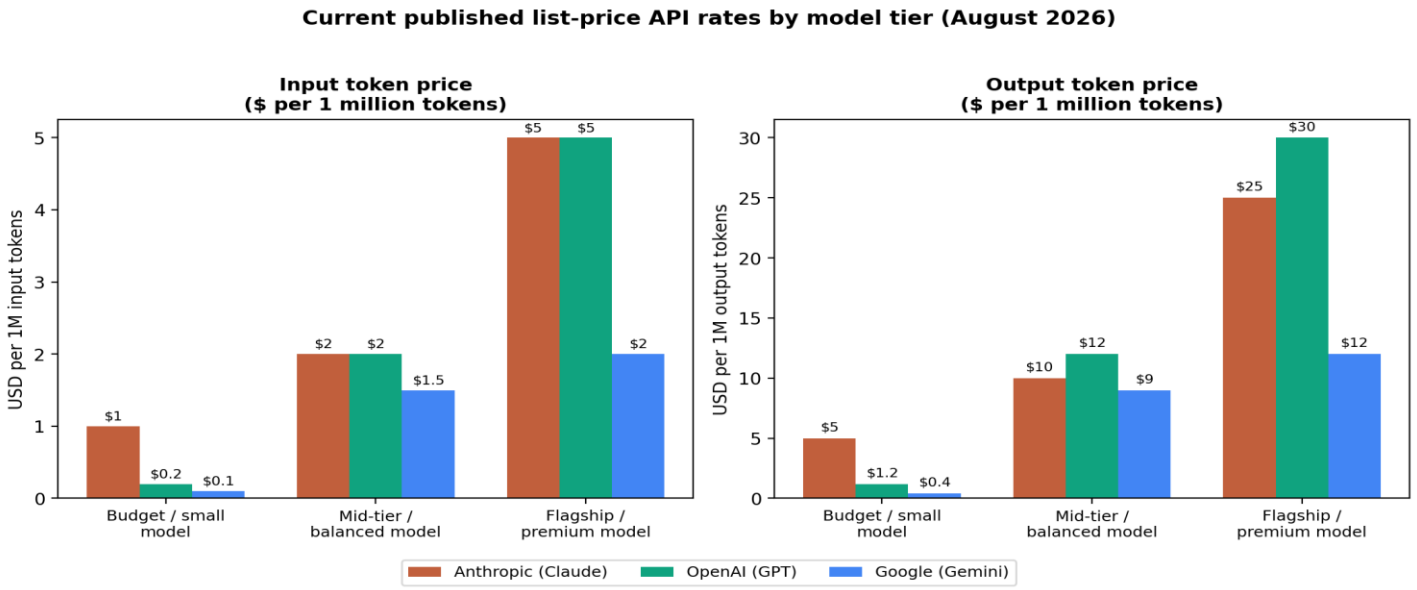
Variable	What it means in plain terms	Typical effect on cost
Which model you use	Vendors sell several models at once: a cheap/fast one, a balanced one, and a premium "flagship."	Can differ 10 – 50x between the cheapest and most capable model from the same vendor.
Length of what you send in (input tokens)	Your prompt, plus any documents, prior messages, or instructions attached to it.	Scales directly with size: pasting a 50-page contract costs much more than a one-line question.
Length of what comes back (output tokens)	The model's written response.	Output is priced 4 – 6x higher than input on every major vendor. A verbose answer costs noticeably more than a terse one.
"Thinking"/reasoning effort	Newer models can work through a problem step-by-step before answering. That internal reasoning is billed as output, even though the user doesn't see all of it.	Harder, more open-ended questions can quietly consume far more tokens than the visible answer suggests.
Agent/tool-use steps	An "agent" that searches, reads documents, and takes several steps to complete one task makes multiple model calls per task, not one.	Multistep, autonomous workflows can use several times more tokens than a single question-and-answer exchange.
Volume/enterprise discounts	Heavy users can often negotiate custom rates directly with the vendor.	Not publicly listed: available case-by-case above a certain usage level.

Useful mental model: cost is closer to "pay for what the model reads and writes, multiplied by how capable a model you asked for" than a flat per-question fee. A short, simple question to a budget model can cost a small fraction of a cent. A long document handed to a premium model, with a detailed multipage answer, can cost 100 – 1,000 times more per exchange.

3. How Claude, GPT, and Gemini Price Today

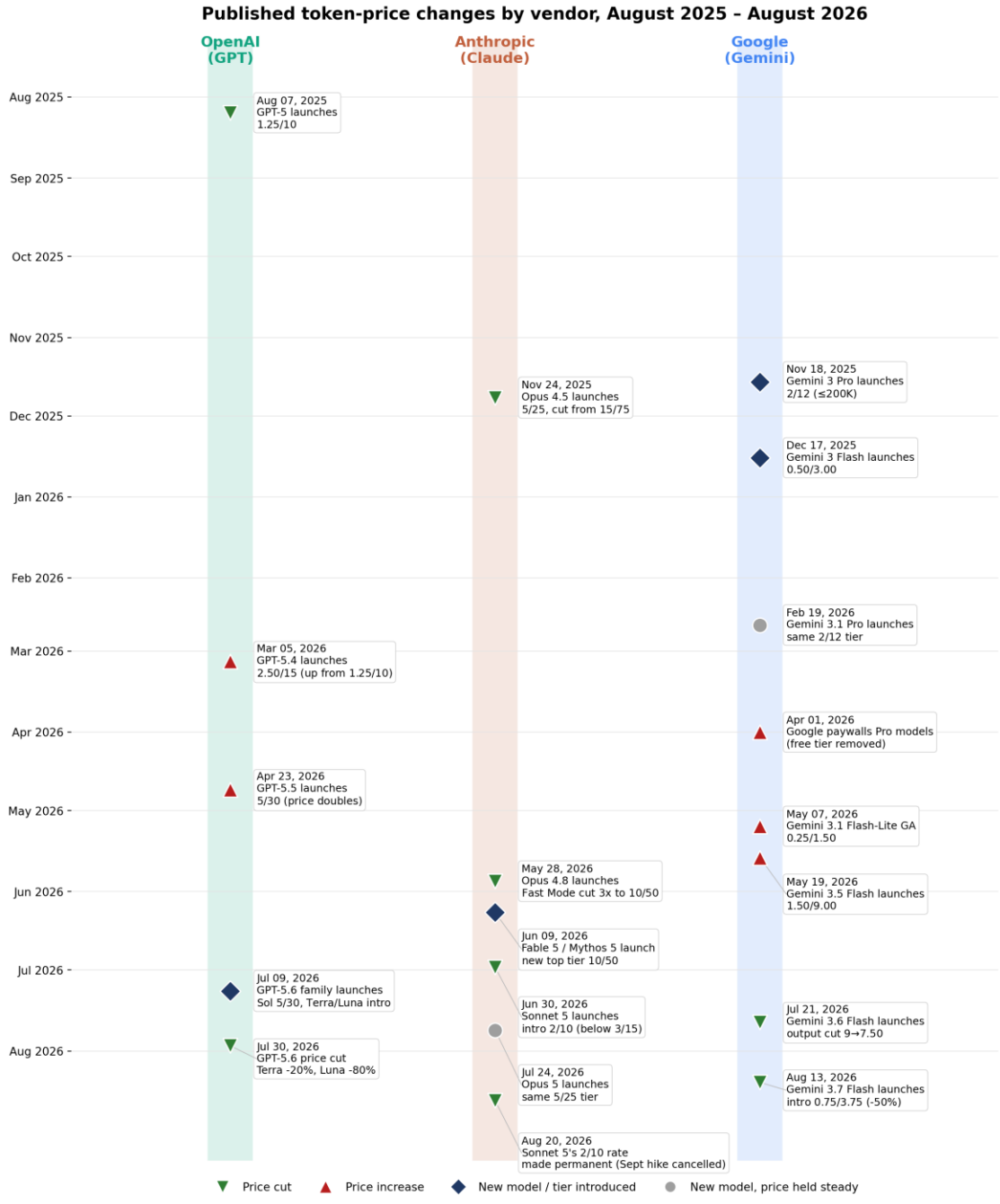
All three major vendors use the same basic structure: separate input and output rates per million tokens, with model tiers ranging from cheap-and-fast to premium-and-capable. The table below shows current, directly verified list prices as of August 20, 2026. If you are reading this guide after August, pricing is likely to change based upon the AI vendors described below.

Model (tier)	Input $/1M tokens	Output $/1M tokens	Output vs. input
Anthropic – Claude Haiku 4.5 (budget)	$1.00	$5.00	5x
Anthropic – Claude Sonnet 5 (balanced)	$2.00	$10.00	5x
Anthropic – Claude Opus 5 (flagship)	$5.00	$25.00	5x
Anthropic – Claude Fable 5 (top/Mythos tier)	$10.00	$50.00	5x
OpenAI – GPT-5.6 Luna (budget)	$0.20	$1.20	6x
OpenAI – GPT-5.6 Terra (balanced)	$2.00	$12.00	6x
OpenAI – GPT-5.6 Sol (flagship)	$5.00	$30.00	6x
Google – Gemini 2.5 Flash-Lite (budget)	$0.10	$0.40	4x
Google – Gemini 3.5 Flash (balanced)	$1.50	$9.00	6x
Google – Gemini 3.1 Pro, ≤200K context (flagship)	$2.00	$12.00	6x



Output always costs more than input (roughly four to six times), and the gap between the cheapest and most expensive models within the same vendor can be 25 times, or more.

Pricing has changed nearly every month over the past year: at least one major vendor adjusted rates in ten of the last 13 months. That volatility makes contract flexibility important.



4. Subscription vs. Metered Pricing: How the Market Is Evolving

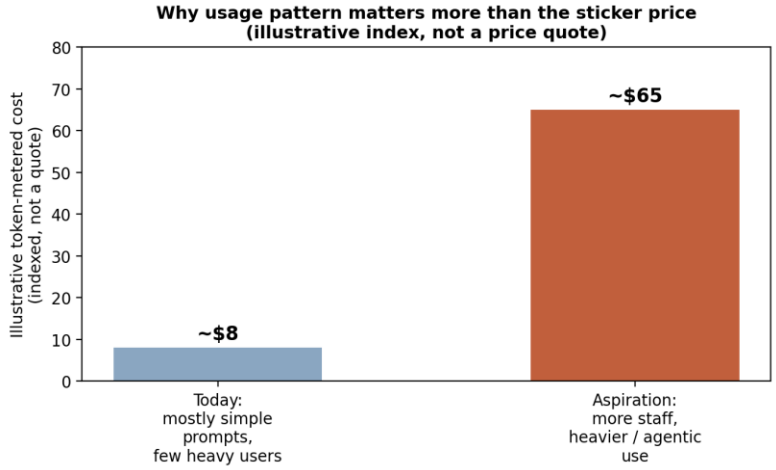
It helps to distinguish three commercial models that AI vendors commonly use:

- **Flat subscription:** a fixed monthly or per-seat fee for unlimited or generously capped use. This is simple to budget for and can work well when actual usage per customer is low and fairly uniform. The vendor is betting that most customers will not use enough to make the flat fee unprofitable;
- **Metered/usage-based:** you pay for what you actually consume, typically denominated in tokens, per-document, or per-query, sometimes with a base platform fee layered on top; and
- **Hybrid:** a base subscription that includes a usage allowance, with metered overage charges beyond that; increasingly common as vendors' own AI costs become a larger share of what they spend to serve you.

As AI usage grows more complex, the market is shifting from flat subscriptions toward usage-based models.

5. What This Means for Your Organization

- Track which teams, matters, and task types drive AI usage.
- Expect a shift toward usage-based pricing as work becomes more complex and compute-intensive.
- Require transparency on model usage, included allowances, overages, and price-change rights.



6. Look Beyond the Vendor's Price: How Are You Actually Using AI?

Understanding what your AI vendor charges is only part of the equation. Your organization should also consider whether it is deploying AI efficiently and whether your commercial model reflects how the technology is actually being used.

- Does the task require an LLM at all? Some repetitive or deterministic tasks may be handled more efficiently through search, rules-based automation, templates, or conventional software. Even when an LLM is appropriate, consider whether a lower-cost model can achieve the desired business outcome rather than defaulting to a premium model.
- Does your licensing model reflect actual usage? A flat per-seat price may be attractive because it is predictable, but not every licensed user necessarily uses AI the same way. Understanding how many users are active and whether usage is broadly distributed or concentrated among a relatively small number of sophisticated users can help determine whether seat-based, tiered, hybrid, or usage-based pricing makes sense for your organization.
- How will increased adoption change the economics? Today's AI usage may consist primarily of simple prompts and document reviews. As your organization moves toward multistep workflows and agents, consumption patterns may change significantly. Consider whether the pricing structure that works today will remain attractive as AI becomes more deeply embedded in everyday work.

7. Key Questions to Ask Your AI Legal Technology Vendor

The following questions can help any legal department better understand its AI vendor's pricing structure and plan for cost management.

Questions to Ask Your AI Legal Technology Vendor
Is the new pricing metered per token, a tiered subscription with usage caps, or a hybrid (base fee plus overage)?
Which underlying model(s) power the product, and can your organization choose a cheaper model for routine tasks and reserve a premium model for complex matters?
Do you pass through the vendor's list-price token rates 1:1, or do you apply a markup, a blended rate, or your own bundled rate?
What happens to price if usage shifts from mostly simple lookups toward longer, more complex agentic workflows?
Can your organization receive a usage dashboard or itemized report showing token consumption by matter, user, or task type, so spend can be monitored as it changes?