Executive Primer

The True Cost of AI Tools

What firms should know about token pricing.

June 2026

12601297769 | 2026 August 03

Table of Contents

Introduction

Many AI tools used in legal practice today, including Harvey and Legora, are no longer priced like ordinary software. Instead of paying a fixed annual licence fee, firms increasingly pay based on how much the AI is used, measured in units called "tokens". The headline price quoted by a provider is almost always only part of the real cost. As usage grows, particularly as these tools take on bigger and more complex tasks, the bill can grow with it, sometimes sharply and with little warning.

Providers do not always provide a clear, indicative estimate of what costs will look like once a tool is used at scale across a firm, and the way usage translates into cost is not always explained in plain terms. A firm can sign up to a tool on the strength of a modest pilot price, only to find that the same tool, used the same way, costs substantially more once it is rolled out to every team and embedded in daily work, because the underlying mechanics of how cost builds up were never made clear at the outset.

Used well, these tools deliver real value. But this is a reason to understand, in plain terms, what is driving the cost, and to think carefully about how a tool is adopted, and by whom it is built and managed. This report sets out that detail, and the practical means by which the risk can be controlled.

What is a Token, and Why Does it Matter?

A token is the basic unit an AI model uses to process text. As a rough guide, one token is approximately four characters, or about three-quarters of a word. A 1000-word document typically equates to roughly 1,300 – 1,500 tokens.

Every interaction with an AI model involves two types of tokens, both of which are charged, usually at different rates:

- **Input Tokens**, the text sent to the model: instructions, retrieved documents, prior conversation history, and any other supporting context.
- **Output Tokens**, the text the model generates in response.

Output tokens are consistently more expensive than input tokens across all major providers, typically by a factor of three to six times, because generating text is more computationally demanding than reading it. A growing number of advanced "reasoning" models also generate internal "thinking" tokens before producing a final answer. These are billed as output even though the client never sees them, and they can be a significant hidden cost driver in complex tasks.

The AI Pricing Model

Traditional software is priced like a subscription: a fixed fee, regardless of how much it is used. AI tools work differently, because the provider itself is charged behind the scenes every time the AI does any work, by the company that built the underlying model (such as Anthropic or OpenAI). The more the tool is used, the more it costs the provider to run, and that cost is generally passed on to the firm, whether as a metered fee, a usage cap with charges for going over it, or some mix of subscription plus usage.

Consequently, the price a firm is quoted at the outset is often not the price it will pay once the tool is rolled out across teams and embedded in everyday work. Two firms using the same tool for the same general purpose can end up with very different bills, depending purely on how heavily and how the tool is used. This difference is rarely visible at the point of signing.

This is becoming more pronounced as tools move beyond simple Q&A into agentic work where the AI carries out multistep tasks largely on its own, such as researching, drafting, and checking its own output with limited human input. These tasks can use far more tokens than a single question and answer, because the AI is, in effect, having an extended conversation with itself behind the scenes to get the job done. Firms are often not shown this detail in advance, and it is rarely reflected in early pilot pricing.

AI Cost Drivers

In practice, a handful of things tend to push usage, and therefore cost, higher than firms expect, often without anyone deliberately doing anything wrong:

- How much material the AI has to read. Every document, clause, or precedent it pulls in to answer a question adds to the cost, so a thorough answer can cost several times more than a quick one even though both look the same to the user.
- How long and detailed the answer is. Longer, more polished responses cost more to generate than short ones, and this is the most expensive part of the bill.
- How complex the task is. Harder reasoning tasks can cost substantially more per query, and some of that cost comes from the AI's internal "working out," which the user never actually sees but is still charged for.
- How many people are using it, and how. As adoption spreads across a firm, small individual costs add up quickly, and usage is very difficult to forecast accurately in advance.

None of this is necessarily disclosed clearly upfront. A provider's published price per query or per user tells a firm very little about what its actual bill will look like once the tool is in daily use across multiple teams. Because these tools are largely a black box from the firm's perspective, built, hosted, and controlled entirely by the provider, a firm has little ability to see, question, or influence any of these drivers directly.

Managing Costs With the Right Approach

None of this means costs are bound to spiral, or that AI tools should be approached with alarm. It means they need to be adopted deliberately and with clear intent, in the same way a firm would manage any other significant and variable cost, not switched on and left to run on their own. There is now an established set of techniques for keeping AI costs under control, and the extent to which a firm can use them depends heavily on how that AI capability is built and who controls it.

1. Sending the right amount of information, not everything available (context trimming). Much of the unnecessary cost in AI tools comes from feeding the system more background material than a task needs. Done well, the AI is given only the most relevant passages for the task at hand, rather than entire documents or libraries "just in case," often cutting the volume processed, and the cost, by a significant margin.

2. Reusing what has already been processed (prompt caching). Where the same background material, house style, or instructions are used repeatedly, a well-built system will reuse that earlier processing rather than paying to redo it every single time. On repeat work, this is consistently one of the largest available cost savings.

3. Matching the tool to the task (model orchestration). Not every task requires the most powerful, most expensive version of the AI. Simple, routine work can often be handled by a lighter, cheaper model, with the most capable and costly version reserved for genuinely complex matters, directing spend to where it earns its keep.

4. Well-designed prompts and processes (prompt engineering and chunking strategy). How a task is set up and instructed to the AI has a direct bearing on cost, not just quality. Thoughtfully designed, well-tested ways of using the tool consistently cost less, and produce more reliable results, than ad hoc, lawyer-by-lawyer use.

5. Visibility before the event, not surprise after it (usage monitoring and cost governance). Tracking and forecasting usage, by team, practice group, or matter, and putting budgets or alerts in place in advance, allows a firm to spot a problem early and attribute spend to real outcomes, rather than discovering it for the first time on an invoice.

This is why the question of how a firm acquires its AI capability matters so much. A third-party tool such as Harvey or Legora is, from the firm's side, a closed system: the firm cannot see how it retrieves information, cannot tune how it uses cheaper models for simpler work, and cannot apply most of the techniques above. They sit inside the code the firm does not control and the provider has limited commercial incentive to optimise.

By contrast, a solution built with and for the firm, its own AI tools and agents, developed in partnership with a provider who designs for cost discipline from day one, allows every one of these techniques to be built in from the outset, tested, and refined as usage grows. Cost control becomes a feature of the build. This is also where a partner who understands both the technology and the firm's own data, workflows, and risk appetite adds the most value: not simply implementing a tool but engineering how it is used so that cost and value move together, deliberately, by design.

As AI becomes embedded in everyday legal work, this kind of deliberate, well-governed approach is fast becoming a genuine point of differentiation, and, for firms operating at scale, increasingly a necessity rather than a nice-to-have.

Taking Control

There are already examples of firms whose AI costs grew rapidly and unpredictably after rollout, particularly where the tool's success was judged by how much it was used, rather than by the value it delivered. Heavy use of a tool is not the same as that tool paying for itself.

The core issue is transparency. The true cost of these tools is often not knowable from the pricing a firm is shown at the outset, and providers do not always volunteer what that cost looks like once usage scales firmwide. Token consumption should be treated as a genuine, ongoing cost of running the tool, on a par with the licence fee itself. The firms best placed to avoid that outcome are likely to be those that retain genuine visibility and control over how their AI capability is built and run, rather than those relying entirely on a third party's word.

This primer is intended as a general, plain-English briefing and does not reflect the specific terms of any individual provider or contract. Commentary reflects publicly available market information as of June 2026 and is subject to change.

epiqglobal.com