Why AI gains aren't compounding at large law firms

Written by Sasha Perrin

Faster tasks, same timelines

AI has compressed the early stages of legal work. Drafting cycles are shorter. Research turns up faster. Analysis that used to take a full afternoon can be generated in minutes. If you look at individual tasks, the improvement is obvious.

If you look at how matters actually move through the firm, the picture gets muddier. In many large firms, matters are starting faster without finishing faster. A draft might be produced in 20 minutes, then sit for days while someone confirms whether it reflects the current negotiated position, checks it against assumptions in the financial model, and reconciles it with direction the client gave over email last Tuesday. The production time dropped. The time between "draft complete" and "ready to use" went up.

That pattern repeats across the workflow. Research surfaces more results, which means more time sorting signal from noise. Faster turnaround on memos creates more iterations, which means more moments where someone has to confirm that everyone is still working from the same set of facts. The effort doesn't disappear. It shows up later, in different forms, often carried by more senior people.

Gains at the point of production are real. But in many firms, the time saved is reappearing downstream as review, coordination, and rework.

This isn't a failing of the tools. It's a consequence of speeding up one part of a system while leaving the rest unchanged. The tools produce output. The organization still has to absorb it.

Where the friction actually lives

Senior leaders at large firms can usually point to the friction without much prompting. The problem isn't that the firm lacks systems or data. Every large firm has document management platforms, billing and ERP systems, matter management tools, email, messaging, knowledge repositories. Each one captures a real and valid piece of the matter.

What none of them captures is the full picture. The state of a matter, at any given moment, is spread across all of these systems, and no single one reflects the whole. So when a lawyer needs to understand where a deal actually stands, or what happened on a litigation matter while they were staffed on something else, they have to pull the picture together themselves. They check the DMS for the latest draft, scan email for client direction, review the billing data for scope changes, ask a colleague what was decided on the last call. This happens constantly. It's baked into how work moves.

The result is that a significant portion of professional effort goes toward assembling a working understanding of the matter rather than advancing it. Handoffs take longer than they should because the receiving lawyer has to rebuild context from scratch. Review cycles expand because the reviewer needs to verify not just the quality of the work but whether it aligns with the current position. Decisions get deferred while someone confirms what is actually current across three different systems.

A firm can have access to every piece of information about a matter and still lack a coherent understanding of where that matter stands. Access and coherence are different problems.

Firms have operated this way for years, and it worked well enough when the pace allowed it. Experienced professionals carried the state of matters in their heads, pieced things together across systems, and absorbed the overhead as a normal part of the work. But that compensation depends on having enough time and enough continuity of staffing to make it work. As pace increases and matters move faster, the overhead of keeping everything aligned scales with it. At a certain point, the firm is generating work faster than it can maintain a shared understanding of that work.

What happens when you add speed to a fragmented system

AI makes the fragmentation problem worse, not better. That sounds counterintuitive, since the tools themselves are genuinely useful. But the effect of accelerating output in an environment where systems don't share state is that every downstream dependency gets hit more often.

When a team can produce three draft versions of a document in the time it used to take to produce one, the review and alignment process doesn't get three times faster along with it. The reviewer still has to check each version against the current deal terms, the financial assumptions, and whatever direction came in over the last 48 hours. If any of those inputs live in different systems (and they do), the reviewer is spending real time pulling context together before they can even begin evaluating the substance.

The same dynamic plays out in decision-making. More available output means more options, more data points, more potential paths forward. But the context needed to evaluate those options hasn't gotten any easier to assemble. If anything, the faster pace means there's less time to do it. So decisions either slow down while someone builds the picture, or they get made against an incomplete one.

Under lower-speed conditions, firms compensate for fragmented context through expertise and effort. As speed increases, that compensation breaks down.

Most firms can recognize this dynamic when it's described. The signals are familiar: review cycles that expand even as production times shrink, meetings that exist mainly to re-sync participants on where a matter actually stands, recurring questions about which version of a document is current. These aren't isolated inefficiencies. They're symptoms of an environment where the speed of work has outpaced the ability to keep it coherent.

Additional tooling doesn't fix this, because each new tool adds another place where a version of the truth lives. Integration layers help with access, but access and alignment are different things. You can make it easier to find a document without making it any easier to know whether that document reflects the current position of the deal.

The architectural question most firms haven't asked yet

Most large firms today sit somewhere between fragmented systems and partial integration. They've invested in making information easier to find, and that investment has paid off. But the state of the matter still resets at key handoff points, and coherence still depends on the people closest to the work. Faster tools, better search, more integrations: all of these help at the margins, but none of them change the underlying condition. In most firms, the state of a matter is something people carry and reassemble, not something the environment maintains.

Changing that is an architectural problem. It requires defining how the matter is represented across systems, establishing authority when sources disagree, governing how state changes over time, and making the resulting view usable at the point of action. With the right framework you can identify exactly where context is breaking down, understand what progression looks like from fragmented to operationalized, and make the governance and design decisions that determine whether AI investment compounds or plateaus.

Free ebook
The full model: From diagnosis to design

The ebook includes the tools to move from observation to action:

✓ A four-stage maturity model for assessing how context behaves in your firm

✓ Four design decisions, with diagnostic tests, that shape the operating model

✓ A matter trace exercise you can run against a live matter this week

✓ The structural model for treating context as operating infrastructure

---

About Clio

Clio is the global leader in legal AI, powering legal work across the profession, from law firms of every size to corporate legal teams, government agencies, and courts. Its Intelligent Legal Work Platform unites the business and practice of law, connecting the people, information, and processes behind legal work. Clio's AI brings the context together with legal sources drawn from Clio Library, a growing collection of more than one billion legal documents. Over 400,000 legal professionals in more than 130 countries rely on Clio to run their legal world. Learn more at www.clio.com.