The Thin Firm and the Quiet Court -- Where the People Go in the 2030s: An Essay on The Future of Legal Work and Machine Justice

By Yvonne @ Legal InnovAI

ABSTRACT

Machines already perform legal work at scale, so the real questions about the future of law are no longer about capability but about people: where they fit when some work no longer needs them, and what becomes of accountability when judgment itself can be automated. This essay traces legal practice two years out and ten years out—the supervised firm, the thin firm, the quiet court—and argues that the most consequential shift is in accountability, which AI does not abolish but unbundles into transparency and the bearing of consequence. It then makes an uncomfortable case: because human justice is covertly biased and its opacity is part of how that bias survives, a transparent, auditable automated system could be fairer than the one it replaces. It closes with the demanding conditions such a system would have to meet—and the human choices that decide whether we build it. Also, the Appendix explores automation readiness for ten types of matters.

> *"The opacity we have prized in human judgment has been hiding its bias—and a transparent machine, built right, could deliver a justice more consistent, more inspectable, and more correctable than anything a covertly biased human system has ever managed."*

By 2026, vendors at Legalweek were routinely demonstrating AI agents that take a single litigation hold notice and—without further instruction—identify the relevant custodians, map the data sources, draft the preservation letters, and schedule the collection in minutes. Major legal-research and e-discovery platforms shipped this functionality in the first quarter.

This scene is where we'll begin, because this author believes that the interesting questions about the future of law are no longer about capability. Machines already do legal work at scale, and the curve is steepening. By early 2026, more than nine in ten legal professionals reported using at least one AI tool in daily work, per the Wolters

Kluwer 2026 Future Ready Lawyer Survey, and adoption of general-purpose generative AI among lawyers had more than doubled in a single year. Major legal research vendors shipped autonomous, multi-step agents in the first quarter. Courts have started piloting the same technology: a Los Angeles County program launched in early 2026 gave a handful of civil judges access to software—a tool called Learned Hand— that distills hundreds of pages of motions and drafts tentative rulings in the judge's own style, with the judges required to review and edit before adopting anything.

Now, some really interesting questions worth asking are about the role of people in the era of AI. If the work can be done by something that doesn't bill, tire, or forget, where do the people go? And—perhaps some harder questions that this essay builds toward—if accountability and even judgment can be partly automated, is that necessarily a loss? Or could a justice system that leans less heavily on human discretion than it used to actually be a *fairer* one than the one we have now? The answers require looking at three things in turn: the near future of legal practice, the structure of accountability itself, and the deeply uncomfortable possibility that the opacity of human justice has been hiding more injustice than the transparency of machine justice would create.

What Law Actually Is

> *Automation makes the mechanical layer of law nearly free, so much of the value migrates to accountability and judgment.*

Law is two things braided together: the mechanical application of rules to facts, and the allocation of risk, responsibility, and legitimacy among people. Automation collapses the cost of the first toward zero while leaving the second largely intact — and a major story of the next decade is what happens to the second when the first becomes essentially free.

To see where people fit, it helps to be precise about what law is, because it is two things braided tightly together, and it's easy to fail to separate them.

Law is, in one part, the mechanical application of rules to facts. Read the document, find the clause, compare it to the standard, flag the risk, cite the authority, draft the response, file by the deadline. This is most of the visible labor of legal practice, and it is enormously automatable. It has structure, precedent, and verifiable outputs.

The second part is the allocation of risk, responsibility, and legitimacy among people. When a lawyer signs a document, they are not merely producing text; they are standing behind it. When a court issues a ruling, it is not merely computing an answer; it is exercising authority that people have agreed to accept. When a client in

the worst week of their life sits across from counsel, they are buying judgment, advocacy, and someone who they believe will be accountable for the outcome.

Much of what follows comes from one fact: automation collapses the cost of the first thing toward zero while leaving the second largely intact. As the mechanical layer becomes nearly free, value, jobs, and meaning migrate to the layer of accountability and judgment.

And here, my friends, is where this article differs from much of what is being said in the industry about judgment not being automatable.

The rest of this essay is in part an argument with what this author believes is an overconfident claim. The truth is probably a gradient, and where the line falls is perhaps the most important design question of the coming decade in shaping the law and legal services.

Prediction for Two Years Out: The Supervised System

> *By 2028 the work changes inside roles—lawyers become verifiers, the billable-hour pyramid cracks, and routine courts are automating.*

By 2028 the work changes inside the roles before reshaping institutions: associates spend their day analyzing and checking rather than gathering, the firm pyramid quietly cracks, and high-volume low-stakes courts begin handling routine disputes algorithmically. Humans remain everywhere — but the reasons they remain narrow sharply to one thing: accountability.

Picture a competent firm—and a competent court—in 2028. The change is real but not yet structural; it has happened inside roles before reshaping institutions.

The associate's day has inverted. Where a junior lawyer once spent most hours gathering—pulling documents, summarizing depositions, running first-pass research—they now spend most hours analyzing, checking, and deciding. One industry observer called this an "80/20 reversal": lawyers spending most of their time on analysis rather than the information-gathering that used to consume it. The phrasing is Robert J. Couture's, recorded in the Harvard Law School Center on the Legal Profession's 2025 study of AI's effect on firm business models: "AI may cause the '80/20 inversion'; 80 percent of time was spent collecting information, and 20 percent was strategic analysis and implications." Agents handle the contract review end-to-end and surface the three clauses that matter; the lawyer's job is to know those are the right three, and to own the consequences if they are not.

This makes verification the central activity, and it is harder than it sounds. Some embarrassing failures making the news of this era are not about agents doing nothing; they are agents doing something confidently wrong. By late May 2026,

Damien Charlotin's AI Hallucination Cases Databasehad catalogued 1,497 court decisions worldwide in which AI-fabricated authorities had turned up in legal filings, up from about 712 at the end of 2025. The count, which Bloomberg Law characterized as "metastasizing," is itself a floor: it captures only decisions where a court explicitly addressed the hallucination, and undercounts by several different mechanisms. Checking the machine becomes a core professional skill, taught explicitly, and the lawyers worth the most are the ones who can look at a fluent, fully-formed work product and reliably find the one place it went wrong. Reading critically becomes more valuable than writing from scratch. The model side is moving too, in ways the two-year horizon should not ignore: Anthropic's Claude Opus 4.8 (released 28 May 2026) is the first frontier model marketed primarily on honesty rather than capability—"more likely to flag uncertainties about its work and less likely to make unsupported claims," per the company's release notes—and Google has positioned its Sec-Gemini work and its newly launched AI Threat Defense (with Mandiant and Wiz) as a cybersecurity counterweight to the same agentic systems lawyers are starting to rely on. The implications of both are explored in "The Technical Floor" later in this essay.

The most common economic model of hourly law firms starts to crack here. The traditional firm runs on a pyramid: a few partners leverage many associates, billing junior hours at a multiple of cost. That pyramid is a machine for converting human labor into profit. When the labor at the bottom is replaced by agents that cost almost nothing per task, the arithmetic of the billable hour stops working. A firm's pricing migrates from time to value, and the firms that move first gain an edge while the others quietly bleed margin.

In the courts, the same two-year horizon shows automation entering at the edges and the bottom. High-volume, low-stakes disputes are already being handled by software in ways most people never notice: online dispute resolution systems have for years resolved tens of millions of e-commerce disagreements annually, the overwhelming majority without a human ever touching them. Estonia has piloted algorithmic handling of small claims under a set threshold, with human judges hearing appeals. Taiwan has tested systems that draft complete rulings—facts, reasoning, citations, and a proposed verdict—for routine categories like driving-under-influence cases, which a judge then reviews and may issue with or without changes. China's internet courts in Hangzhou, Beijing, and Guangzhou use AI to move e-commerce disputes through to judgment in minutes, and a court in Shenzhen has integrated a large language model trained on a vast corpus of Chinese legal text directly into the drafting of judicial reasoning.

The reassuring part of the two-year picture is that humans remain visibly everywhere—they supervise, decide, sign, appear, and review. The unsettling part is

that the *reasons* they remain have narrowed sharply, and most of those reasons points at the same destination: accountability.

Ten Years Out: The Thin Firm and the Quiet Court

> *By the 2030s a dozen people do what a hundred did, and managing a firm becomes governing machines rather than people.*

By the mid-2030s, what changed inside roles has reshaped institutions: a firm of a dozen handles work that a hundred did, the management of legal service delivery becomes the governance of systems, and the courts reserve human judges for appeals, novelty, and mercy. The structural consequence is that leadership shifts from people-management to risk governance.

Extend the line. By the mid-2030s, changes that lived inside roles have reshaped institutions, and a genuinely automated legal practice is a business model rather than a thought experiment.

Imagine a firm handling the legal needs of a few hundred corporate clients with perhaps a dozen people. There is no associate tier—not a thin one, none. The work is performed by fleets of specialized agents: a diligence agent, a drafting agent, a regulatory-monitoring agent, an orchestrator coordinating them, each a hyper-specialized product rather than a general assistant, because the market has fractured into dozens of narrow tools that each do one kind of legal work extremely well. They run continuously. They do not have a Friday. A matter that once took a team three weeks resolves overnight.

So what are the dozen humans doing? A few are principals—lawyers who hold the bar license, carry the insurance, and stand as the accountable party to clients, courts, and regulators. A few are legal engineers who build the workflows, set the guardrails, audit the outputs, and intervene when an agent meets something genuinely novel. A few manage relationships, because clients facing existential risk still want a human who knows their business. And one or two do the work that has no template at all: arguing a question of first impression, persuading a regulator, finding the structure no agent would propose because no agent has ever seen one.

The structural consequence is the one that matters most: if there are hardly any people doing the legal work, the management of legal service delivery becomes something entirely different. A traditional managing partner manages people doing law—recruiting, training, motivating, allocating, building the culture that retains them. This author, whose expertise is in the business management of law firms, is no longer performing the majority of the human-centric work her role required pre-2030.

(See on CounselCommons.com, for example, some of the automation tools designed by Legal InnovAI that are relevant for human capital management at law firms today but will not be as relevant for the people-lite law firms of the 2030s).

In the thin firm there are almost no such people. Management becomes the governance of systems: deciding which agents to deploy, how to verify their work, where to set the boundaries of autonomy, how to handle the cases the system flags as beyond its competence, and how to maintain the accountability chain so that when something goes wrong there is always a human who can be held responsible. Leadership shifts from people-management and rainmaking toward risk governance, quality assurance, and systems design. The firm stops being primarily an organization of professionals and becomes an accountability structure wrapped around a fleet of capable machines—a thin human shell whose job is to be trustworthy, answerable, and able to handle exceptions.

This creates a paradox the profession will spend the decade fighting about. The senior judgment the thin firm depends on was historically forged by years of doing the junior work the thin firm has eliminated. They learned to spot the clause that mattered by reviewing ten thousand contracts as a young associate. If no one reviews contracts the same way anymore, where does the next generation of judgment come from? A profession that automates away its own apprenticeship has to deliberately rebuild the path to mastery, or it runs out of the expert humans that are still necessary in this new system.

The courtroom version of the thin firm is the quiet court: most disputes resolved by systems, with human judges reserved for appeals, novel questions, and the cases where stakes or ambiguity demand discretion. Whether that future is a dystopia or the largest expansion of access to justice in history depends almost entirely on how we answer one question, which the rest of this essay takes up directly: what happens to accountability when the work is done by a machine?

How Accountability Changes

> *AI doesn't abolish accountability; it unbundles transparency from consequence-bearing—and a perfect audit trail still needs a human to hold the bag.*

Accountability is two things fused: the legibility of a decision, and the bearing of its consequence. The human system has been bad at both while disguising the legibility problem behind the comforting presence of a named human. What AI does is not reduce accountability so much as unbundle it — and the genuine danger to engineer away is legible unaccountability, where the log is perfect but no one answers.

Let's start with a careful distinction, because so many confused arguments about AI and responsibility conflate two different things.

An audit trail answers *what happened and why*. Accountability answers *who bears it*. A complete log can show that a system deviated from standard at a particular step, which model ran, which retrieval failed, the entire chain of reasoning. That is traceability, and it is a genuine good. But the log itself has no assets, no license, no standing. You cannot sue it, sanction it, or make it compensate the person it harmed. An audit trail can *establish* fault. It cannot *bear* fault. Fault has to land on something that can carry it.

The reason this matters is that accountability is mostly forward-looking. The reason a lawyer checks a citation is that *they* lose if it is wrong—their license, their reputation, their malpractice premium. That stake produces care in advance. Strip out any answerable subject and the audit trail merely documents the accident with precision after the fact. It improves the autopsy, not the prevention. A bridge can have a flawless safety record and we still hold the engineer accountable, because when the record fails, "who answers" is not answered by past performance.

So the chain of responsibility in an automated system can be long and heavily automated in the middle, but it must terminate in a node that can suffer consequences—a person, or a legal entity ultimately backed by people and capital. Audit trails make that node well-informed and traceable. They do not remove the need for it, because consequences only bite on things that can be made to lose something.

Pay attention now, because here is the turn and it is an important move in the essay. The instinct to treat the human node as the precious, irreplaceable thing rests on a flattering assumption about how accountability currently works. This author believes that assumption is largely false.

The accountability of the present legal system is, to a significant degree, ceremonial. There is a name on the document, but the actual decision is a black box. Why did the judge rule that way? Why did the associate miss the issue? Why did the prosecutor offer that plea rather than another? The real reasoning lived in a human head and is undocumented and now lost. Malpractice is hard to prove precisely because you usually cannot reconstruct what the lawyer actually did or thought. Judicial opinions are written after the decision and may rationalize rather than reveal it. And the consequences rarely bite: lawyers are seldom sued for negligence, and judges almost never face anything for a poor call. The system gives us a *target* for blame while hiding the *basis* for it, and then mostly declines to follow through.

This reveals that accountability is two things fused together: the *legibility* of a decision—can we see what happened and why, audit it, detect patterns, contest a specific step—and the *bearing* of its consequence. The human system has been bad at both while disguising the legibility problem behind the comforting presence of a

named human. What AI and audit trails do is not reduce accountability so much as *unbundle* it. They pull transparency sharply upward and force the bearing-of-consequence out into the open, where it has to be consciously assigned rather than assumed.

That unbundling is mostly healthy, but it carries a specific danger: transparency can become a *substitute* for accountability rather than its completion. The failure mode reads, "the process was followed, here is the complete log proving it, therefore no one is to blame." A perfect record of a harm gets used to diffuse responsibility rather than locate it. That is legible *un*accountability—the harmed person ends up with flawless documentation and no remedy—and it is arguably the most plausible dystopia in this whole story. Transparency is necessary and the current system badly lacks it, but transparency does not convert itself into someone answering and someone being made whole. That conversion is a design choice, and it is the human work we have ahead of us.

The Opacity of Human Justice

The case for a transparent automated system rests on a fact most defenses of human judging would rather not dwell on: an unauditable expert is reputation standing in for inspection, which is the same proxy problem this whole shift is supposed to end.

Now push the argument where it is least comfortable, into the courtroom itself, because the case for machine justice rests on a fact most defenses of human judging would rather not dwell on: human justice is already biased, and its opacity is part of how the bias survives.

The evidence is not subtle. A widely discussed study of parole decisions (Danziger, Levav, & Avnaim-Pesso, “Extraneous factors in judicial decisions,” Proceedings of the National Academy of Sciences, 2011) found that the share of favorable rulings started high at the beginning of a session and declined toward almost nothing as the judges grew tired and hungry before a break, then recovered after they had eaten—though researchers have rightly cautioned that some of this may reflect how cases were scheduled rather than blood sugar alone (see Glöckner, “The irrational hungry judge effect revisited,” Judgment and Decision Making, 2016). The narrower point survives the caveat: extraneous factors that have nothing to do with the law or the facts do indeed move human decisions, and the people making those decisions cannot see it happening.

The racial evidence is graver, and it spans decades. Since 2010 the United States Sentencing Commission has, in a recurring series of analyses, found that Black male

offenders receive longer federal sentences than similarly situated White male offenders after controlling for the legally relevant factors—offense severity, criminal history, and the like. Its 2017 report put the gap at 19.1 percent over fiscal years 2012–2016, and at 20.4 percent once violence in an offender's criminal history was taken into account; its 2023 report, covering 2017–2021, still measured a 13.4 percent gap (U.S. Sentencing Commission, Demographic Differences in Federal Sentencing, 2017 and 2023). The disparity does not disappear even among defendants sentenced within the same guideline range, where the Commission found a 7.9 percent difference—about as close to comparing identical cases as observational data allow. Controlled work points the same way. Examining inmate records with equivalent criminal histories, Blair, Judd, and Chapleau found that within each race, defendants with more pronounced Afrocentric facial features received harsher sentences (Psychological Science, 2004); and Eberhardt and colleagues found that, in cases with White victims, Black defendants whose appearance was rated more stereotypically Black were more likely to be sentenced to death ("Looking Deathworthy," Psychological Science, 2006)—bias operating below the threshold of anyone's awareness, in a system whose participants sincerely believe they are being fair.

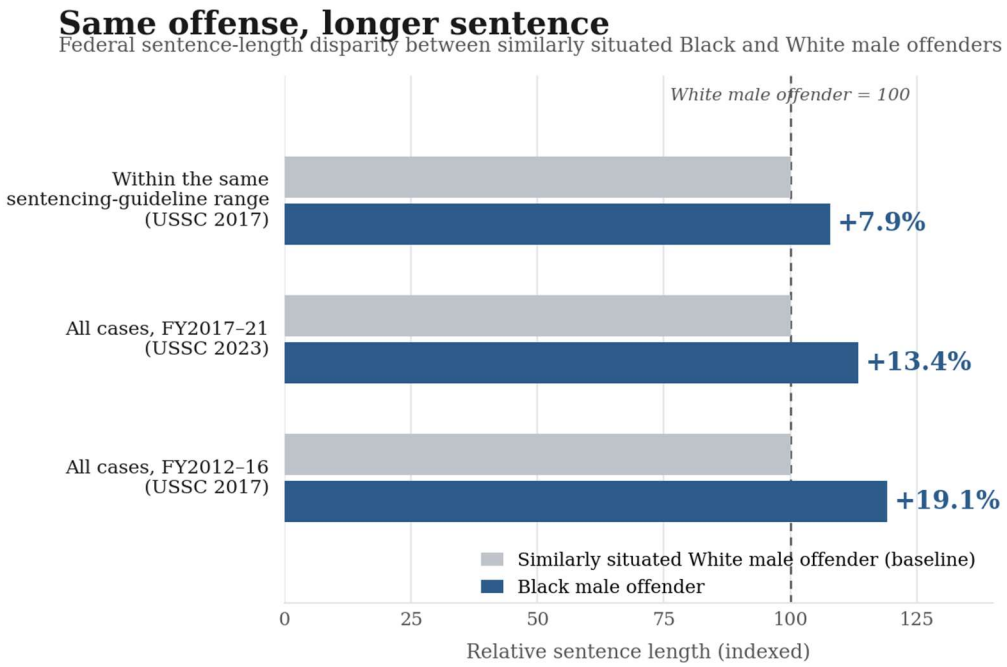


Figure 1. *Federal sentence length for similarly situated Black vs. White male offenders, indexed to a White-male baseline of 100. "Similarly situated" denotes a regression-adjusted comparison that controls for offense severity, criminal history, and other legally relevant factors—not two literally identical cases. The 7.9% figure compares defendants sentenced within the same guideline range. Source: U.S. Sentencing Commission, Demographic Differences in Federal Sentencing (2017; 2023).*

Hold these two pictures side by side. In the human system, the reasoning is invisible, the bias is undetectable at the level of the individual case, and the only "accountability" is a name attached to a decision whose true basis no one can recover or review. You cannot audit ten thousand human sentencing decisions for the influence of a defendant's face, because the influence was never recorded and the judges say they do not believe it occurred. The opacity is not incidental to the injustice. It is what allows the injustice to continue, decade after decade, while everyone involved claim they sincerely believe they are being fair.

That is the setup for the genuinely provocative claim: a transparent automated system, even an imperfect one, could be an improvement over a covertly biased human one—not because the machine is wise, but because the machine is *inspectable*.

Why an Automated Court Might Be Fairer

> *A transparent machine can be fairer than a covertly biased human, because the romantic defense of discretion relied on no one being able to look inside.*

Four legs hold up the case: consistency, auditability at scale, the removal of irrelevant influences, and contestability of the actual reasoning. The case for human discretion has quietly relied on the discretion being unexaminable; transparency is uncomfortable because it would show how unjust the status quo already is.

The argument has several distinct legs, and it is worth laying each out before turning to the obvious objections.

1. The first is consistency. A defining feature of human justice is that the same case decided by two different judges—or by the same judge before and after lunch—can come out differently. This inter-judge variation is itself a form of injustice: like cases are not treated alike. A well-specified automated system applies the same standard to the same facts every time. Consistency is not the same as correctness, but arbitrary variation is a recognized wrong, and removing it is a real gain.

2. The second leg, and the strongest, is auditability at scale. This is the mirror image of the opacity problem. With distributed human decisions you cannot detect systematic bias at the level of reasoning, because the reasoning is never recorded. With an automated system you can run the entire decision population through tests for disparate impact across protected groups, find the patterns, and—crucially—*correct* them. For the first time, bias becomes a measurable, addressable property of the system rather than an invisible

feature of ten thousand private minds. A human judge who unconsciously penalizes a defendant's face cannot be debugged. A system can be.

3. The third leg is the removal of irrelevant influences. The machine is not hungry, tired, annoyed by the lawyer's tone, or primed by the previous case. Whatever its flaws, they are not the flaws of a depleted human making the twentieth consequential decision of a long afternoon.

4. The fourth leg is contestability of the actual reasoning. When a human denies your claim, the stated reasons may not be the real ones, and you have nothing to argue against but a conclusion. When a transparent system denies your claim, you can in principle see which step drove the result and challenge that specific step. The decision becomes a thing you can take apart rather than a verdict handed down from an opaque authority.

Put together, these legs support a claim that should be stated plainly because it cuts against the romantic defense of human judging: the case for human discretion has quietly relied on the discretion being unexaminable. We trust the human partly *because* we cannot see inside, and what we cannot see has been hiding bias, fatigue, and arbitrariness all along. Transparency is uncomfortable precisely because it would show us how unjust the status quo already is.

The Counterargument

> *But algorithms can launder biased history, standardization can grind away mercy, and a court is not only a decision engine.*

Three serious objections: algorithms can launder bias when trained on biased data; standardization grinds away mercy, proportion, and the capacity to depart from the rule when the rule would produce injustice; and a court is not only a decision engine — the right to be heard is the right to be heard by someone who can be moved.

A serious version of this argument has to meet the strongest objections head-on, and they are strong.

1. The first and most important is that algorithms are not innocent of bias—they can launder it. The cautionary example is the COMPAS risk-assessment tool used in American sentencing and parole decisions, which drew sustained criticism for producing racially skewed risk scores (the criticism originating in Julia Angwin et al., “Machine Bias,” ProPublica, 2016, with technical debate continuing since) and which the courts allowed to be used despite its workings being a proprietary secret. This looks at first like a decisive refutation: the machine was biased too. But read carefully, COMPAS is

actually an argument *for* the thesis, not against it. Its central sins were that it was a *black box*—proprietary, unauditable, its logic shielded from the very defendants it judged—and that it was trained on data reflecting a biased history. It reproduced human injustice while wearing the mask of objectivity, and no one could open it to check. The lesson is not "algorithms are biased, so keep humans." It is "opaque algorithms are as bad as opaque humans, and the entire benefit comes from transparency." A secret algorithm combines the worst of both worlds: the inflexibility of code with the unaccountability of a sealed mind.

This points at a deep technical problem: training data. A system that learns from past human decisions inherits their bias by construction. If you train a sentencing model on decades of sentences shaped by racial animus, it will faithfully reproduce that animus and present it as neutral computation. Automation does not wash history clean; it can ossify it. This is the real meaning of the "garbage in, garbage out" warning, and it is why the transparency that makes bias *detectable* is necessary but not sufficient—you also have to be willing to act on what you detect, including refusing to treat the biased past as ground truth.

2. The second objection is about what gets lost when discretion gives way to standardization. Legal scholars have warned that AI adjudication pushes toward what they call "codified justice"—a paradigm that favors standardized rule-application over the discretionary, equitable, case-by-case moral judgment that has been considered the most precious exactly in the hardest areas, like criminal justice (but where it is also biased and broken in other ways still). Mercy, proportion, the recognition that this defendant's circumstances are genuinely different, the capacity to depart from the rule when the rule would produce an injustice—these are probably still features of human judgment until a machine can process enough context to fairly evaluate unique situations. Without that machine, a system optimized for consistency may grind them away. There is also a related worry that machine adjudication, anchored in past data, could freeze the law in place and impede the slow, human evolution by which legal standards change as society's values change. This author believes, however, that a machine given new data can learn and apply new rules more quickly than humans.

3. The third objection is the one that resists every technical fix: a court is not only a decision engine. The right to be heard means the right to be heard by someone who *can be moved*—who can recognize suffering, weigh credibility, perceive remorse or vulnerability, and respond as a person to a person. A system can model these things statistically; it does not experience

them, and many people may reasonably feel that being judged by a machine is itself an injustice, however accurate the output. And there is a legitimacy dimension that is partly empirical: if people do not accept algorithmic verdicts as authoritative, trust erodes, appeals multiply, and the efficiency gains evaporate. The data we have on this is early and pointed in a direction the optimistic case should not avoid: studies of AI deployed as an *assistive* tool — Ecuador's courts, Singapore's Small Claims Tribunal, Estonia's small-claims process — find efficiency and consistency gains, while the closest experimental study of AI *replacing* a judge's decision (Imai and co-authors' randomized evaluation of a pretrial risk-assessment instrument) found that algorithmic substitution made the classification worse, not better. The bottom-tier work courts will keep automating; the moment a system is asked to stand in for the judge rather than help her, the case for it gets harder, and the right to be heard by someone who can be moved becomes a constraint the design has to honor rather than route around.

None of these objections is fatal to the case for machine justice. But together they define the conditions under which it could be true.

What Would Be Required

> *An automated court beats a human one only if it meets hard conditions: open reasoning, standing bias audits, clean data, real appeal, and a human who answers for it.*

Suppose we take seriously the possibility that an automated justice system could be fairer than the human one it replaces—not in the abstract, but in practice. What would actually have to be true? The conditions are worth stating as a standard against which any real system should be measured.

Eight demanding conditions:

1. *Transparency by construction*
2. *Continuous auditing with a duty to correct*
3. *Training data that does not launder injustice*
4. *A preserved right to contest before a human*
5. *A human-bearing accountability node with real remedy*
6. *Genuine rather than theatrical explanation*
7. *Democratic legitimacy over the encoded values*
8. *Deliberate room for equity, mercy, and evolution*

Transparency by construction. The system's reasoning must be open to inspection—by the parties, by reviewers, by the public—not a proprietary secret like COMPAS. A justice system whose logic cannot be examined is illegitimate regardless of its accuracy, because the entire advantage over human opacity is the ability to see inside. This rules out closed commercial black boxes deciding consequential cases, however well they benchmark.

Continuous auditing for disparate impact, with a duty to correct. It is not enough that bias *can* be measured; the system must be audited as a standing practice across protected groups, the results made public, and there must be an enforceable obligation to fix disparities that surface. Transparency that no one acts on is theater. The whole point is to convert the new visibility of bias into its actual reduction.

Training data that does not launder injustice. Because models trained on biased history reproduce it, the design must confront the data problem directly—through curation, through fairness constraints, through a refusal to treat past human decisions as the definition of correct. This is genuinely hard and partly unsolved, and honesty requires admitting that a system cannot be cleaner than the effort put into cleaning its inputs.

A preserved, meaningful right to be heard and to contest. Every person subject to an automated decision must be able to challenge it before a human with real power to override—not a rubber-stamp review, but genuine recourse. The automated layer handles volume and consistency; the human layer handles the cases where standardization would be unjust, where the facts are genuinely contested, or where mercy is owed. The right to appeal to a person who can be moved is not a nostalgic luxury; it is what keeps the system answerable to the humans it governs.

A human-bearing accountability node with remedy attached. Recall the distinction: transparency is not accountability. The system must terminate in someone—an official, an institution—who is answerable for outcomes and against whom a wronged person has a real remedy. The danger to engineer away is legible unaccountability: perfect logs, no one to blame, no one to pay. Someone must hold the bag, and the path from harm to remedy must be short and real.

Explanation that is genuine, not theatrical. A system can produce a fluent rationale that has little to do with how it actually reached its result—the machine equivalent of a judicial opinion that rationalizes a decision made on other grounds. Real contestability requires that the explanation track the actual basis of the decision closely enough that challenging the explanation challenges the decision. An impressive-looking justification for an opaque process is a more convincing fiction, not a truer one.

Democratic legitimacy over the values encoded. Every adjudication embeds value choices—how to weigh competing interests, what counts as a relevant factor, where to set the thresholds. In the human system these choices hide inside discretion. In an automated system they must be made explicit, and once explicit they cannot be left to vendors or engineers. They have to be set through a legitimate public process, because encoding the values of a society into the machinery that judges that society is a political act, and political acts require consent.

Room for equity, mercy, and evolution. The system must be built with deliberate channels for departing from the rule where the rule would produce injustice, and it must not freeze the law. Standards have to remain revisable as values change, which means the humans in the loop are not only exception-handlers but the means by which the law continues to grow. A justice system that cannot change is not just; it is merely consistent.

Notice what these conditions have in common. Every one of them is a place where humans do not disappear but move—from making each decision to designing, auditing, governing, and standing behind the system that makes decisions, and from invisible discretion to explicit, contestable, accountable choice. The automated court does not eliminate the human role in justice. It relocates it to the most consequential point: deciding what justice the machine is for.

What Can Be Automated, and When

> *Forecast by task, not job title: information work is automated now, end-to-end agents and machine adjudication are on the threshold, and accountability resists.*

The honest way to forecast automation is not to ask whether "lawyers" or "judges" will be replaced, but to break their work into tasks and ask which tasks are already commoditized, which sit on the threshold, and which resist automation for reasons that are not merely technical. Legal work is not one thing; it is a bundle, and the bundle is coming apart at different speeds.

Already in production are the information-handling tasks — the parts of law that are search, extraction, comparison, and first-draft generation. Document review and e-discovery, legal research, contract analysis and clause extraction, due-diligence triage, deposition and record summarization, intake and matter triage, and routine document automation are done by machines at scale today; more than nine in ten legal professionals report using at least one AI tool in daily work, and major vendors shipped agentic, multi-step versions of these workflows in the first quarter of the year. On the justice-system side, the equivalent "now" is online dispute resolution: high-volume, low-value civil disputes already resolve through automated negotiation and tribunal pipelines, often without a human ever touching them.

On the threshold are the tasks that stitch those pieces into end-to-end work, and the tasks that begin to exercise judgment under supervision: coordinated agents that carry a whole matter from intake to filing, automated negotiation and algorithmic facilitation that propose settlements, predictive case assessment, continuous regulatory monitoring, and — most consequentially — machine-drafted adjudication, already piloted for routine categories such as the Los Angeles County tentative-ruling tool and Taiwan's draft-judgment systems for high-volume offenses. These are not science fiction; they are this year's pilots, one procurement cycle from production.

What resists automation is what this essay has already named: the accountable node, judgment in genuine novelty, advocacy as one human persuading another, and the value-laden choice of ends. The frontier moves through the first two tiers quickly and stalls at the third — which is exactly why a fully-automated process is best understood not as a robot lawyer or a robot judge, but as a pipeline that automates the commoditized layers, escalates only the residue, and reserves the human for the apex. The diagram below stitches the automatable pieces into one such pipeline for a high-volume civil claim, using a design that already exists in working form.

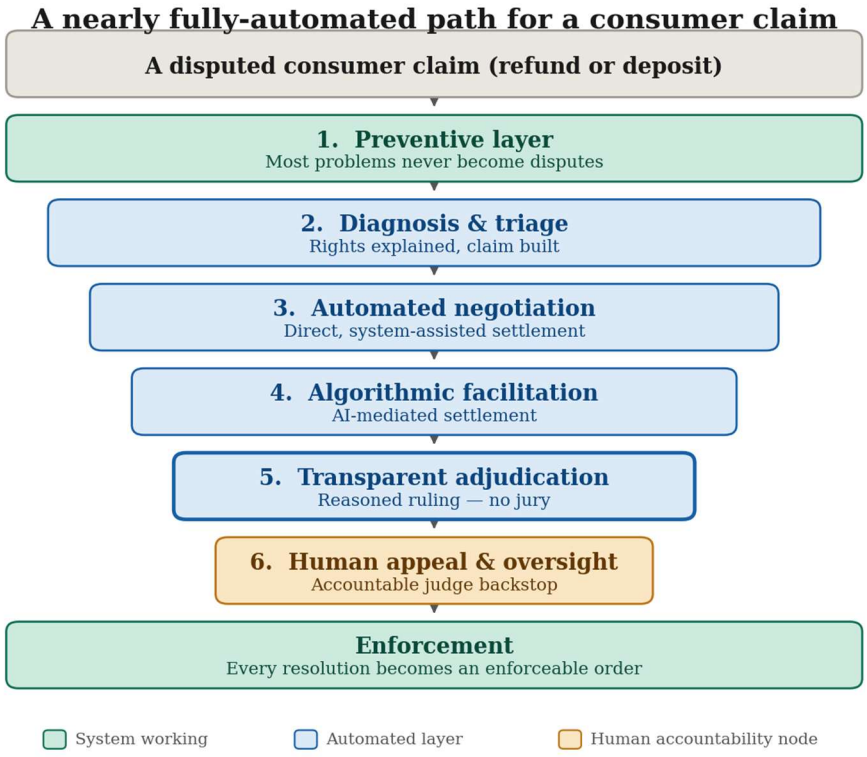


Figure 2. *A nearly fully-automated pipeline for a high-volume civil claim. Most disputes are prevented or resolved in the upper, automatable layers; only the residue reaches adjudication, and every automated ruling is reasoned, auditable, and appealable to an accountable human. The design deliberately contains no*

simulated jury. Model in operation: British Columbia's Civil Resolution Tribunal, a four-stage online tribunal whose decisions are enforceable as court orders.

Read top to bottom, the pipeline inverts the traditional escalation toward trial: prevention removes most problems before they are disputes, automated diagnosis, negotiation, and facilitation resolve most of what remains, and only the residue reaches a transparent adjudication engine whose every ruling is reasoned and appealable. Note what it does not contain — a simulated jury. A jury's legitimacy comes from real peers serving as the conscience of the community and a check on power, functions that a panel of role-playing models cannot supply; so the honest automated end-state replaces it with a transparent reasoned decision, community values set in published and contestable rules rather than improvised by synthetic jurors, and a guaranteed human appeal. This is automation in line with the law as it already is: jury rights attach to people and are routinely waived, and the great bulk of civil disputes already resolve without a jury ever being seated.

The Technical Floor: Honesty, Accuracy, and Cyberdefense

> *None of this is safe without engineering preconditions that did not yet have a serious answer a year ago—models that flag their own uncertainty, pipelines hardened against attack, verifiable provenance for every assertion the system makes, and independent evaluation infrastructure. The first two have moved fast; the third is starting; the fourth is dangerously behind.*

The legal-political conditions rest on engineering preconditions that were optimistic a year ago and are now active competitive frontiers among the labs that matter: calibrated honesty in the models, hardened pipelines underneath them, verifiable provenance of every assertion, and the independent evaluation infrastructure treated in its own section below. The frontier has moved on the first two; on the third, the work is starting; on the fourth, the field is dangerously behind. What remains is whether the legal system insists on benchmarks before lending these tools its authority, or accepts the vendors' word.

The conditions in the prior section assume several things that, until very recently, were optimistic: that a frontier model can be made to reliably admit when it does not know something; that the pipeline carrying its output can be defended against an adversary determined to manipulate it; that every assertion the system makes can be traced back to a source the user can independently inspect; and that the whole stack can be measured against benchmarks the public can trust. Each is a precondition the others rest on. Calibrated honesty is what makes the audit trail worth auditing; cyberdefense is what keeps the chain of custody from being silently corrupted; verifiable provenance is what makes "transparency" mean something more than fluent self-explanation; and independent evaluation is what tells the legal system

whether any of it is working. The encouraging change, in the months running up to the writing of this essay, is that both have become open competitive frontiers among the labs whose models likely could be the substrate of any automated legal system. That shift in posture is moving the "what would be required" list from aspiration to live engineering work.

Consider the honesty side first. On 28 May 2026 Anthropic released Claude Opus 4.8, the first frontier model marketed primarily on calibrated honesty rather than raw capability. The company's own framing—the model is "more likely to flag uncertainties about its work and less likely to make unsupported claims"—is precise about which failure mode it is targeting: not factual error in the abstract, but the confidence with which fluent systems present untrue things as facts. That is the specific failure that has produced the 1,497 sanctioned filings catalogued by Charlotin and counting. The accompanying alignment-science work, including Anthropic's May 2026 "Teaching Claude Why" research direction, treats honesty as something the model is actively reinforced into displaying rather than something it acquires by accident—and which, in early reports from external testers, shows up as the model's greater willingness to say "I am not sure" on questions it would previously have answered with manufactured confidence. None of this is solved. But it is the first time a frontier lab's flagship release has put the honesty axis on the lead line of its marketing, which is itself a market signal: calibrated uncertainty is now a thing customers buy, and labs compete on. For a legal system whose entire complaint against current models is their unwarranted confidence, that is the change that matters most.

The cyberdefense side is the other half of the floor. A reliable justice or legal-services pipeline cannot be silently manipulated, cannot leak privileged work product, and cannot become an attack surface through indirect prompt injection or training-data poisoning. Google's 2026 announcements—Sec-Gemini for analyst workflows, the agentic security-operations center introduced at RSAC 2026, and the AI Threat Defense suite—describe a deliberate move to put defender-side automation on the same footing as the agentic tools attackers are starting to use. The pitch is autonomous discovery, prioritization, and remediation of vulnerabilities at the speed adversaries are now exploiting them. Whatever one thinks of any particular product, the fact that the major-model labs are now competing publicly on cyberdefense indicates that the model is hardened, the pipeline is hardened, the data is hardened has graduated from an afterthought to a commercial frontier. For automated legal services that route privileged communications, client identifying information, and matter strategy through models accessed over the network, that is the precondition that has to be addressable before any of the legal-political conditions in the prior section can stand on it.

Now the third pillar, which is where the work is actively starting rather than well underway: provenance. Calibrated honesty tells you when the model is unsure; it does not tell you what the model is sure on. For automated legal services, the harder demand is that every factual or legal assertion the system produces can be traced, by the user, back to the specific source that supports it—the cited case, the statutory provision, the document in discovery, the line in the contract. Without that, "transparency" collapses into the model fluently explaining itself in English, which is the failure mode the essay has been arguing against from the opening pages. The encouraging movement here is structural rather than headline-grabbing: retrieval-augmented systems are increasingly built with citation strings that resolve to the underlying passage rather than to a model-generated paraphrase; Anthropic, OpenAI, and the legal-research vendors have all shipped variants of "which part of which document supports this sentence" as a first-class output rather than an afterthought; and the W3C and academic groups are converging on machine-verifiable provenance standards for AI-generated content. None of this is solved, and the user-experience design for inspecting provenance is still poor in most products. But the pieces are in place, and a deployment that does not give the user a click-through path from any assertion to its source is now visibly behind, in a way it was not a year ago.

A caveat, before the section closes, is that marketing claims are not measured outcomes. A model marketed as "most honest yet" is still a model whose calibration the legal system has to verify rather than accept on the lab's word; a cyberdefense stack pitched as autonomous still has to be benchmarked against red-team adversaries the public can inspect. The right posture is that the technical floor is becoming achievable, not "achieved," and that the burden of demonstrating each part of it sits firmly on the vendor, not on the public the system would govern. Nevertheless, the right baseline a year ago was that the floor was not even being seriously aimed at by the labs producing the substrate, and that is no longer true. The frontier has moved, and the legal-system conditions of the prior section—transparency, audit, contestation, remedy—now have something underneath them they can plausibly be built on.

The Apprenticeship Problem in Detail

> *Automating away junior work severs the path that produced senior judgment, so the profession must rebuild the ladder on purpose.*

Senior judgment was forged by years of doing junior work that is now being automated. This author believes that the profession will run out of seniors unless it does three concrete things: reverse-engineer the judgment ladder into explicit competencies, teach supervised reading of fluent machine output as a graded skill,

and deliberately expose juniors to the long tail of unusual cases that agents will otherwise suppress.

The earlier sketch raised but did not really answer a paradox the next decade will have to face: senior judgment was forged by years of doing the junior work that automation now eliminates. A profession that automates its own apprenticeship runs out of the seniors it cannot replace. What would have to happen for that not to be the consequence? Three concrete things.

1. The first is the deliberate reverse-engineering of the judgment ladder. Medicine had to do this once it could no longer rely on raw volume of cases to produce mastery: residency restructured around explicit competencies, simulated exposures, and supervised reasoning rather than "see one, do one, teach one" alone. Law's version is overdue. The patterns that juniors used to absorb implicitly—the way a covenant matters in a particular kind of M&A deal, how a specific judge reads a Daubert motion, the tells that distinguish a strong settlement posture from a weak one—have to be identified explicitly, taught in curated sequences, critiqued under supervision, and tested. The 2026 Law Student Pulse Survey from the Thomson Reuters Institute shows that the people on the receiving end of the gap already feel it: 72 % of students name AI literacy as essential, and 74 % simultaneously worry that over-reliance will hollow out their own competencies. They are right to worry. The profession owes them a structured answer.

2. The second is supervised reading as a graded competency. If the new junior skill is critical reading of fluent machine output—recognizing the place a model overstates its support, the place its chain of reasoning drifts off the prompt, the place a citation looks plausible and is not—it has to be taught with the seriousness drafting once was. Most J.D. programs do not yet treat AI-checking as a core competency; bar examinations do not test it directly; firms do not reliably promote based on it. None of those defaults will change without deliberate effort, and each of them has to.

3. The third is deliberate exposure to the long tail. Senior judgment partly comes from having seen rare cases that pattern-match against the current one; the partner who recognizes the unusual posture of a witness, or the term sheet that looks innocuous and is not, is drawing on outliers seen long ago. Agents create the inverse risk: they suppress variation, feed back to associates a smoothed and plausible-but-narrow distribution of past matters, and quietly erase the outliers that built mastery. Counteracting that requires curated case banks, simulation, and supervised exposure under partners who deliberately surface the unusual—deliberate training inputs rather than byproducts of doing the work.

The structural point is that firms which automate junior labor and then under-invest in this rebuilding will see their senior bench dry up over the next decade, and will not realize it has happened until they cannot staff a hard matter. The investment in apprenticeship-by-design is unglamorous, expensive, and slow to repay—exactly the kind firms tend to defer. The ones that don't defer are the ones whose name still means something in 2036.

The Token Economy

> *Inference costs have fallen forty-fold and are now traded like oil—commoditization arriving as a futures curve, with all its discipline and cruelty.*

The cost of a frontier inference token has fallen roughly sixty-fold in three years, and the same tokens are now being financialized — the Shanghai Futures Exchange and the CME are racing to launch contracts that let businesses hedge AI compute the way they hedge oil and electricity. For law, this is the moment Richard Susskind predicted twelve years ago arriving with a vengeance: commoditization not as a slogan but as a futures curve.

Every claim made earlier about cheap, abundant legal labor rests on a hidden assumption: that the unit cost of running a model continues to fall. So far it has, dramatically. Inference for GPT-4-level capability fell from roughly $60 per million output tokens in early 2023 to roughly $1 per million output tokens by May 2026 on the cheapest production-grade tier (Anthropic's Haiku 4.5 and equivalents) — a sixty-fold reduction in three years on that band, driven by mixture-of-experts architectures that activate only a fraction of their parameters per request, better chips, and brutal competition among providers. The catch, which the optimistic story tends to elide, is that the curve is not one curve. The cheap tier (Haiku 4.5 and its peers, at roughly $1 input / $5 output per million tokens) has fallen sixty-fold. The flagship tier of today (Anthropic's Opus 4.8, OpenAI's GPT-5.5, Google's Gemini 3 Pro) prices at $5–$10 input and $25–$50 output per million tokens — cheaper than the flagship of three years ago per unit of capability, but five-to-twenty-five times more expensive in absolute dollars than today's cheap tier. Open-weight pricing sits below even Haiku: DeepSeek's V4-Pro, comparable on many benchmarks to last year's closed flagships, is priced at $0.435 input / $0.87 output per million tokens since its May 2026 permanent discount, roughly an order of magnitude below Haiku and twenty-to-thirty times below the proprietary flagships. Three curves, not one, and the spread between them has been widening, not narrowing, over the past year. If the cheap-tier curve continues, the access-to-justice arithmetic works out for any legal work that the cheap tier can handle competently: a legal-aid brief that consumed $40 of compute on yesterday's flagship consumes a few cents on today's cheap tier or essentially nothing on a self-hosted open-weight model. The harder

question, and the one this section will turn to, is which legal work belongs on which curve.

But the curve is not guaranteed to continue, and three things are happening at once that complicate it. The first is demand. As capability has gotten cheaper, the volume of inference has exploded. Cheaper per token has meant more tokens, not lower bills — and the additional tokens are disproportionately flagship-tier tokens consumed by reasoning-heavy and agentic workloads, where the per-matter cost has, in some firms, gone up rather than down.

The second is that the cost structure underneath inference is not infinitely compressible. The cost of a token is ultimately the cost of the electricity, water, semiconductor capacity, and capital that produce it. As models grow for reasoning-heavy work, and as agentic systems take many more inference steps per matter than single-shot queries did, the physical cost per matter can rise even as cost per token falls. A long multi-step agentic legal workflow is not the same product as a short query or running a tightly-designed skill— and it almost always runs on the flagship tier, where the bill scales with both the number of steps and the price per step.

The third — and this is the news of the past few days, as I write — is that the AI compute layer is being financialized as a commodity. On 28 May 2026, Reuters reported that the Shanghai Futures Exchange is in early design of AI-token futures contracts, while CME Group and the Intercontinental Exchange in the United States are preparing GPU-compute futures to do the same job on the hardware side. The proposed design treats tokens the way electricity futures treat kilowatt-hours: a non-storable commodity whose price businesses and infrastructure operators need to hedge against. Academic groundwork already exists; an arXiv paper from March 2026 proposes a Standard Inference Token unit. Larry Fink has begun publicly framing compute as a new asset class. The legal profession is watching this slowly because it does not feel like a legal story, but it is exactly the story: when the input to a service has a published forward curve, the people who buy that service gain new pricing power, and the people who sell it have to choose between hedging exposure and pricing it through.

This is the moment Richard Susskind predicted, more or less precisely, in Tomorrow's Lawyers twelve years ago: a world of internet-based global businesses, online document production, commoditized service, legal process outsourcing, and web-based simulated practice. The piece he could not have known the shape of (because, who could?) was the unit on which commoditization would clear. It turns out to be the token, and the token is now headed for the same trading screen as crude oil. Susskind was correct that the law would, eventually, be sold the way commodities are sold, with all the discipline and all the cruelty that implies. His more recent work

extends the prediction into the AI era explicitly; the futures-market news of last week is the receipt.

What does this mean concretely for the essay's main argument? Three things. First, the cheap-help promise depends on the curve, which is no longer purely deflationary. Firms and legal-aid organizations planning two years out should not bet on continued steep price declines; they should plan for a flat or jagged curve where capability rises but per-matter cost does not always fall in step. Second, hedging will become a real corporate-finance line item for any organization whose service-delivery cost is dominated by inference — and the largest in-house legal departments will be the first to demand it from their outside counsel. Third, the firms that survive the transition with their margins intact will be the ones that have engineered their workflows so a given matter can run on the cheapest model that meets a quality bar, and can fall back to the next tier when that one fails. The honest version of this engineering challenge is that some legal work — the high-stakes, the genuinely novel, the cases where a small reasoning error compounds into a large client harm — will not pivot to the cheap tier without losing the quality bar, and the firms doing that work will have to absorb flagship-tier prices into their cost structure rather than wish them away. Which brings us to the pivots, with the bifurcation in view.

Pivots If Tokens Stay Expensive

> *If compute stays dear, the answer is smarter architecture—model cascades, retrieval, and small specialized models—not surrender.*

If inference does not get steadily cheaper, the answer is not surrender. It is a different architecture — model cascades that route easy questions to small models and only the hard ones to frontier, retrieval against curated public-law corpora rather than running everything through an LLM, and distilled task-specific models that capture most of frontier capability for a fraction of the cost. A sixth pivot, treated at length in the "Open Weights and the Sovereignty Question" section below, is to move inference off third-party APIs entirely and onto open-weight models running on hardware the firm controls — which converts the per-token operating cost into a flat amortized capital cost, with the added benefit that the firm's data never leaves its perimeter.

Suppose, contrary to the optimistic scenario, the token-cost curve flattens or rises. The right response is not to abandon the access-to-justice promise; it is to stop assuming the substrate that delivers it must be a frontier model.

1. The most direct pivot is the model cascade. In any defined legal workflow — discovery review, contract analysis, intake triage — the great majority of decisions are easy in expectation; the work is choosing which of the few hard ones to escalate. A cascade routes every input through a small, cheap, fast

model first; only items that model flags as hard, ambiguous, or high-stakes climb to a mid-tier model; only those the mid-tier flags climb to the frontier. The economics inverts the naïve 'everything on the frontier' architecture: ninety to ninety-five percent of the work runs on tokens that cost a fraction of frontier prices, and the firm pays frontier rates only on the residual. The same logic generalizes to verification, where a smaller model can check the output of a more expensive one — the inversion the essay's earlier sections describe, now made financial.

2. The second pivot is distillation. A specific, well-defined legal task — privilege review, redaction, citation-checking, M&A clause classification — does not need the broad reasoning of a frontier general model to run well. A frontier model can be used once, expensively, to generate a high-quality training set (this is the type of product that's ideally suited for the CounselCommons.com marketplace); a smaller model is then fine-tuned on that set to perform the specific task at roughly equivalent quality at a fraction of the per-inference cost. The legal-tech vendors who have understood this are already shipping task-specific models priced well below frontier APIs and running them locally where data sensitivity demands it.

3. The third pivot is retrieval against curated public-law corpora — often called RAG, or retrieval-augmented generation. Most legal questions do not require the model to know everything; they require it to know one statute, one regulation, one case, and to reason carefully about how that source applies to a specific set of facts. A small model with excellent retrieval, against a well-curated corpus of public law and the firm's own playbooks, can beat a frontier model running on its ambient knowledge alone, because the retrieved source grounds the reasoning and shortens the model's path to the right answer. The cost differential is order-of-magnitude; quality on real legal tasks is often better, not worse, provided the corpus and the retrieval pipeline are built with the care drafting once received. (See Legal InnovAI's article "How to Improve Results of Retrieval-Augmented Generation (RAG)")

4. The fourth pivot is using older or cheaper model generations for verification rather than generation. A model that was state-of-the-art twelve months ago is now a small fraction of its price at API; if its job is to read a fluent output — and where the verification model is an older closed-source generation, there is a practical risk that the provider will retire it, breaking the workflow; the durable answer is to use an open-weight self-hosted model in the verification slot, since once the firm has the weights, no third party can deprecate the capability out from under them and find the place it went wrong — the central new junior skill the essay has been describing — it does not need to be as

smart as the model that produced the output. It only needs to be different enough to catch what the first model missed.

5. The fifth pivot is caching. Legal work has unusually high overlap across matters. The same statutory definitions, the same common clauses, the same governing-law boilerplate, the same case-law extractions appear repeatedly across files. A serious caching layer — at the level of retrieved passages, classified clauses, and reusable extractions — can collapse the marginal token cost of a new matter to near zero on everything the firm has already seen, and isolate inference cost to genuinely new questions.

What none of these pivots solve, on their own, is the fundamental problem that firms with the deepest pockets will run frontier models on frontier prices because they can. Unless the cost of frontier models for small firms, non-profits, and legal aid continues to be subsidized without strings attached, the access-to-justice version of automated law depends on the legal-aid and small-firm tier engineering its workflows around small models, retrieval, and distillation. The qualifier is that an organization serving the A2J tier at high volume and low margin per matter can build economics that genuinely rival the boutique-firm on a frontier-tier setup; the math depends on volume, not per-matter willingness to pay, and the rest of this essay treats the size of that potential market explicitly. It is unglamorous and exactly the kind of investment that, like apprenticeship-by-design, often gets deferred until the bills force it.

A coda the bifurcated opener of this section asked for: the pivots in this section answer the question of how to keep the cheap-tier curve as the dominant cost structure for the great majority of legal work, and most legal work that follows clear procedure can be made to fit. The pivots do not, on their own, answer the question of what to do about the work that does not. High-stakes commercial litigation, novel constitutional questions, complex transactional work where a small reasoning error compounds into eight-figure client exposure — these are matters where the quality bar lives at the flagship tier, and engineering them down to the cheap tier with cascades and distillation may simply lose the quality the client is paying for. The honest answer for that tier is not engineering it away. It is admitting that flagship-tier inference is a real and growing cost of doing serious legal work in 2026, pricing it transparently into the matter, and treating frontier-tier prudence (smaller agentic loops, tighter prompts, more aggressive caching, escalation only where escalation is warranted) as a craft skill on par with the cheap-tier engineering the rest of this section described. The bifurcation the opener flagged is not a temporary anomaly to be optimized out. It is the cost structure the next decade of legal work will run on, and the firms that recognize that honestly — building two operating models, one for

the cheap tier and one for the flagship tier, with deliberate routing between them — will outperform the ones that pretend a single curve still tells the whole story.

Evaluations as the New Bottleneck

> *Knowing whether a legal automation is good is harder than building it, and evaluation becomes the scarce, decisive skill.*

Knowing whether a legal automation is good is harder than building one. The benchmarks that exist today are useful and partial. The question of whether evaluation itself can be automated has the same shape as the apprenticeship problem — if it can, juniors lose their training ground; if it cannot, the profession does not scale.

A theme runs through every section so far: the lawyer's new job is to check the machine. But 'check' is doing a lot of work in that sentence, and the question of how to evaluate a legal automation at scale is one of the under-discussed bottlenecks of the next five years.

What exists today is a handful of public benchmarks. LegalBench, released in 2023 by a Stanford-led consortium, is the most widely cited; it covers more than 150 tasks across issue-spotting, rule application, and statutory interpretation. Vendors quote their LegalBench scores the way coders quote SWE-Bench. Benchmarks of this kind are useful — they let two models be compared on the same problem set, and a model that does poorly on LegalBench will almost certainly do poorly in production. But the relationship in the other direction is much weaker. A model that scores well on a public benchmark may still fail on the specific distribution of matters a particular firm sees, because benchmarks measure what is measurable and the profession's real work is full of edge cases that are not.

The deeper problem is that for high-stakes legal work, the right standard is not 'answers the benchmark correctly' but 'answers correctly when it knows, refuses when it does not, and flags its uncertainty in ways the supervising lawyer can act on.' Calibrated honesty, in other words — the property Anthropic's Opus 4.8 is being marketed on — is harder to benchmark than raw accuracy, because the right answer to many legal questions is some version of 'this is not knowable from the materials available,' and that answer is rewarded by good lawyers and punished by naïve evaluators.

There is a second-order question that follows. Could evaluation itself be automated — could a model grade another model's legal work? In principle, yes; in practice, the same problem returns. A grader model evaluating a producer model is a system whose error may correlate with the producer's error where both models share training data and inductive biases. Independent grading requires diversity: a

different model family, ideally trained on different data, ideally with a different architecture. The legal-tech firms that have grasped this run their gradings on a model deliberately different from the one producing the work, and combine the automated grade with sampled human review.

The connection back to the apprenticeship problem is uncomfortable. If grading is automated, the work that was supposed to teach juniors how to read critically — finding the place a fluent output went wrong — is itself done by another machine, and the apprenticeship gap widens. If grading is not automated, the verification bottleneck swallows every efficiency gain the system was supposed to produce, because each matter still ends with a human lawyer reading the entire output. The honest answer is that both will happen in tension: some grading will be automated with a sampled human-review layer attached, and the right design is not 'no humans in the loop' but 'humans deliberately deployed where the system's grading is least trustworthy.'

What the profession actually needs — and does not yet have — is a public, continuously updated, jurisdiction-aware benchmark suite that covers the long tail of real legal work, includes calibrated-uncertainty tests, and is run by an institution whose independence from the vendors it measures is durable. The closest things to that institutional role today are a handful of academic centers (Stanford's Legal Design Lab and CodeX; the Suffolk Legal Innovation and Technology Lab) and a small number of access-to-justice organizations doing the unglamorous work of testing tools against real client matters. They need to be much bigger (is this the future work of the government and can it be incorruptible?). Until they are, a buyer evaluating a legal-AI vendor is essentially relying on the vendor's own internal evals, which is the same governance problem the essay began with.

Open Weights and the Sovereignty Question

> *Open-weight models now trail the frontier by a hair, letting firms, legal aid, and courts run capable systems they understand and can control.*

In 2026, the gap between the best open-weight models and the closed frontier is measured in single benchmark points rather than generations. A mid-sized firm, a legal-aid organization, or a state court system can now run a competitive model on hardware it controls. Whether the profession takes that option, or accepts dependence on a handful of frontier labs, is the choice that decides whether legal AI ends up looking like open infrastructure or like the financial-data terminal market. (This author suggests you buy some serious hardware to preserve optionality and sovereignty.)

A note on terminology before the section continues, since the three labels matter differently. Closed-weight models (Anthropic's Claude, OpenAI's GPT, Google's

Gemini) are accessed only through their providers' APIs; the weights stay on the vendor's servers, and the user controls neither the model nor the data path. Open-weight models (DeepSeek, Llama, Qwen, Mistral, Gemma) release their trained parameters under a license that permits inference and usually fine-tuning, so a firm can download and run them on its own hardware—but the training data and the full reproduction pipeline are typically not released. Open-source AI in the strict sense, as the Open Source Initiative defined it in October 2024, requires the training data, training code, and weights all to be open; vanishingly few production models meet that bar (Allen AI's OLMo and EleutherAI's Pythia are the cleanest examples), and the models this section discusses are open-weight, not open-source. For the legal-system sovereignty argument that follows, what matters is the inference-layer freedom that open-weight licensing provides—the ability to run the model on hardware you control, audit its outputs, and not have the capability deprecated out from under you. The deeper open-source ideal of full scientific reproducibility is a separate and more demanding standard the field has barely begun to meet.

The whole argument so far depends on the substrate. If the only credible models are produced by Anthropic, OpenAI, and Google, then the legal profession's nominal independence from any vendor is a fiction, and the conditions of the 'what would be required' section — transparency, audit, contestation, remedy — collapse into 'whatever the three labs decide to allow.' That risk is now genuinely contestable, because the open-weights frontier has moved more in the past year than in the prior three combined.

As of May 2026, DeepSeek V4 Pro under an MIT license. Kimi K2.6 from Moonshot. The closed frontier still leads, modestly, on the very hardest reasoning benchmarks; but for the great majority of legal workloads — contract review, statute lookup, deposition summarization, discovery classification, drafting under supervision — the open-weights frontier is at parity or close to it, at a fraction of the running cost, with the weights physically present on the firm's hardware.

Four possible concrete deployment patterns describe the realistic options for the profession.

Tier one: the small firm or legal-aid organization. A mid-tier consumer GPU (an Nvidia RTX 4090 or 5090, around $1,500–$2,500) or an Apple Mac Studio with 192GB unified memory ($4,000–$8,000) runs a 30-billion-active-parameter open-weight model — Qwen 3.6-35B, Mistral Small 4, Gemma 4 31B — at conversational speeds. The capital cost amortizes against API spend within months for any organization with sustained inference volume. The model runs offline; client matters never leave the building. The trade is some capability headroom for full data sovereignty and a flat capital cost in place of a metered operating cost. For legal-aid

organizations whose client base is by definition vulnerable and whose clients' confidentiality is the entire product, this tier deserves some real thought.

Tier two: the mid-sized firm or in-house legal department. A small inference rack of four to eight enterprise GPUs (Nvidia H200, B200, or AMD MI300X-class accelerators, around $25,000–$40,000 each) or a managed sovereign-cloud deployment with the firm's own model weights and isolated tenancy. This runs a frontier-class open-weight mixture-of-experts model — DeepSeek V4 Pro, Qwen 3.6 235B, Kimi K2.6 — at production volumes. The capital is real but bounded; the operating cost is predictable; the legal-privilege story is much simpler than 'trust the API vendor's data-handling commitments.' The frontier-API option remains available as a fallback for the hardest matters, but it stops being the default for everything.

Tier three: the state bar association, court system, or sovereign government. A larger inference cluster, ideally jointly procured by multiple courts or bars to share capital cost, with deliberate independence from any single vendor's release cadence and a contractual commitment that the model weights and the corpus it retrieves against remain under public control. This is the option that converts the 'transparency by construction' condition from aspirational to operational, because the public can in principle audit a system whose substrate is open in a way it can never audit a hosted API.

Tier four: the large law firm. A dedicated on-premise inference cluster funded as IT capital expenditure, running a frontier-class open-weight mixture-of-experts model fine-tuned on the firm's own playbooks, precedents, and matter history; supplemented by client-specific or matter-type-specific small models (M&A, securities, complex commercial litigation, regulatory) distilled from the frontier model on the firm's curated work product. The capital is real — mid-seven to low-eight figures — but it is the same order of magnitude as the firm's annual legal-research and matter-management spend today, and it converts an operating expense into an asset. The advantage is structural: the firm controls its model the way it controls its associate training program, the work product never leaves the perimeter, and the practice-specific small models are competitive moats other firms cannot rent. There is also a shared option for firms unwilling to absorb the full capital cost on their own: a multi-firm consortium platform on the model of the existing legal-research platforms but with shared open-weight inference infrastructure and practice-area-specific small models that consortium members contribute training data toward and share access to. Such a platform would compete with closed-frontier vendors on the cost dimension and with closed-frontier products on the data-sovereignty dimension; it does not exist today and arguably should.

The 'infrastructure capture' risk that runs underneath this question is older than AI. The legal-research market consolidated into a Westlaw/LexisNexis duopoly four decades ago, and the profession has paid rent ever since. Financial-market data lives behind Bloomberg terminals because once an industry becomes dependent, the price of leaving rises faster than the price of staying. If every legal-AI product in 2030 is a thin wrapper on three frontier models, the bar associations agents may, two decades on, be writing essays about why their function depends on infrastructure they do not control and cannot replace (assuming essays are still a thing in two decades; they may not be). The window during which it is cheap to avoid that future is the window we are in now.

Mark A. Cohen has argued in Legal Mosaic and across his Forbes columns, for nearly two decades, that the legal industry's transformation has been delayed not by the absence of tools but by the legal profession itself — that change arrives only when buyers force it. The profession is relying more on technology than ever, and the open-weights frontier of 2026 is the first technology in this transformation that gives buyers — corporate legal departments, state bars, legal-aid organizations, sovereign courts — the option to refuse dependence on the closed-frontier vendors rather than negotiate from a position of weakness. Whether the profession takes that option is exactly the kind of choice Cohen has spent two decades pointing at: technical, unsexy, and decisive of who holds the lever in the decade after this one.

Some concrete recommendations, then, for firms to explore.

For a small firm or legal-aid organization: deploy a 30-billion-active-parameter Apache-2.0 or MIT-licensed open-weight model on local hardware as the default workhorse, with a frontier-API fallback for the hardest matters.

For a mid-sized firm: stand up an in-house inference cluster running a frontier-class open-weight mixture-of-experts model, with the cluster controlled by the firm's IT function and the model weights formally subject to the same retention discipline as case files.

For a state bar or court system: lead a public procurement of shared inference infrastructure, on terms that lock the substrate to open weights and the corpus to public legal materials, with vendor-provided closed-frontier models permitted only as an audited supplementary layer. None of this is hypothetical; every component exists today. What is missing is the institutional will to assemble it before the alternative becomes the default.

The Justice Gap, Disaggregated

> *The 92% justice gap is not one bottleneck but six, and the gap with the most people behind it has the fewest tools.*

Access to justice is not one problem but a sequence of failures at six distinct points. Knowing which gap an intervention closes is the difference between organizing the field by org chart and organizing it by leverage. The two facts that emerge from the disaggregation are uncomfortable: the gap with the most people behind it is the one with the fewest tools, and technology can make justice more inspectable without making it more equal.

It is tempting, when listing the organizations doing access-to-justice work, to organize them by what kind of institution each one is: vendor-independent nonprofit here, consumer-facing service there, federal funder over there, university research lab in the corner. That framing is the org chart of the A2J world dressed up as strategy. It tells the reader the field contains nonprofits, government, and universities, which the reader already knew. A more useful frame might ask what each intervention actually does — and the way to do that is to disaggregate the justice gap into the distinct failures that produce it.

The civil justice gap, on close inspection, is a chain of six failures. A person experiences a legal problem and never recognizes it as legal. They recognize it but have no map of what to do. They have the map but cannot navigate the procedure. They navigate the procedure but need substantive advice that the unauthorized-practice-of-law rules forbid a non-lawyer from giving them. They get the advice but face a represented opponent with a structural power advantage. And in some matters the substantive law itself offers them no real remedy at all. Each link is a different problem with different leverage. An intervention only matters insofar as it unblocks a specific one. The Legal Services Corporation's 2022 Justice Gap Study, which underwrites most numbers in this field, found that ninety-two percent of low-income Americans with civil legal problems received inadequate or no help in the prior year. That ninety-two percent is not one bottleneck; it is six.

Figure 3. *The civil justice gap as a chain of six failures: a person can fall out at any link, so only a fraction of those with a legal problem reach a just result. Source: Legal Services Corporation, Justice Gap Study (2022), which found that ninety-two percent of low-income Americans with civil legal problems received inadequate or no help.*

Gap 1 — Issue recognition. A wage statement is short by sixty dollars; an eviction notice arrives without the right notice period; a denial-of-benefits letter cites a regulation that does not apply. The person experiencing it does not see a legal problem. They see a frustration, a setback, or their own fault. Tools at this gap are thin, because the intervention has to find people who do not yet know they are looking. Adjacent infrastructure exists — 211 referral lines, BenefitsCheckUp and mRelief for benefits screening — but legal-issue spotting at population scale is mostly aspirational. The most ambitious frontier is passive monitoring: a model that reads a tenant's letter, a worker's pay stub, or a patient's medical bill and flags it as a possible legal matter without the person having to ask. This is where the largest justice gap sits and where the field has the fewest tools.

Gap 2 — Triage and pathways. Once a person recognizes the problem, they need to know what to do, in what order, by when. LawDroid's LawAnswers AI, launched nationwide in September 2025, is the most visible new entrant: a free public-facing assistant designed specifically to triage civil legal problems and route users toward the right next step. Illinois Legal Aid Online has done this at state scale for two

decades. Pro Bono Net's LawHelp Interactive provides the underlying interview infrastructure to many of the others. This is the strongest link in the chain right now.

Gap 3 — Procedural navigation. The person knows what to file; they need help filing it. Upsolve, founded inside Harvard Law School's Access to Justice Lab in 2016, has helped roughly eighteen thousand low-income families relieve more than eight hundred sixty million dollars in debt through a free Chapter 7 bankruptcy filing tool. JustFix.nyc does the equivalent for eviction defense in New York; Hello Divorce does it for uncontested divorce. The infrastructure under most of this — Documate, Afterpattern, and others — is increasingly AI-augmented. This is the most mature category and the one where the next layer of capability has the clearest payback.

Gap 4 — Substantive advice and the UPL wall. The line between legal information, which non-lawyers may provide, and legal advice, which only licensed lawyers may, varies by state. The line is unevenly drawn and is the largest regulatory obstacle to A2J automation. Perhaps the most-watched test of UPL doctrine in a generation, Upsolve v. James, took the opposite arc. Upsolve and a South Bronx pastor sued the New York Attorney General in 2022, arguing that the state's UPL statutes, as applied to a program training nonlawyer "Justice Advocates" to help low-income defendants respond to debt-collection lawsuits, violated the First Amendment. The Southern District of New York granted a preliminary injunction in May 2022 on strict-scrutiny grounds, and the access-to-justice field treated the decision as a doctrinal breakthrough. The Second Circuit vacated and remanded in September 2025, holding that UPL statutes are content-neutral regulations of speech, subject to intermediate rather than strict scrutiny. On remand, the SDNY dismissed the case in March 2026, finding that the statute survives intermediate scrutiny. Upsolve has petitioned the Supreme Court for review. Read carefully, the case is now a cautionary marker rather than a beachhead: the most ambitious constitutional theory for opening gap 4 to nonlawyer advice has, at least for now, lost. The implication is that the field's actual path through gap 4 runs through state-level regulatory reform—Utah's sandbox, Arizona's alternative-business-structure rule—rather than through a federal constitutional shortcut. This is a regulatory fight more than a technology fight, and it is where most of the policy energy in the field belongs.

Gap 5 — Power inside the proceeding. A person who has identified the problem, navigated to the right court, filed the right papers, and received the right advice can still lose to a represented opponent who knows the judge, can outlast them in motions, and holds the asset. Pro-se litigant tools are early and few — Courtroom5 is one of the rare mature efforts, and AI-augmented preparation for asylum hearings and eviction defense is just beginning. This is the thinnest link in the chain. It is also the link where the essay's larger argument hits a real boundary.

The case the essay is making is that automation, built right, can make justice more inspectable than human justice has ever been. Inspectability is necessary; it is not sufficient. A pro-se litigant with a perfectly transparent procedural-navigation tool and the right substantive advice can still lose to a represented opponent who has more money, more time, and a senior judge's ear. That is not a failure of the argument; it is the argument's honest limit: a transparent, well-built automated system can make procedure more navigable without changing who has the money, the time, and the standing to use it well.

Gap 6 — The underlying law. Some justice gaps are not gaps in access; they are gaps in the law itself. The remedy the person needs does not exist, or exists in theory and not in practice. This is outside the scope of A2J technology and properly the territory of legislative advocacy and impact litigation. It is named here only so the reader knows the prior five do not exhaust the problem.

Two observations fall out of organizing the field this way. The first is that the leverage is currently bunched at gaps 2 and 3 — the links where the field is most mature and the next round of AI improvements will pay back fastest. The second is that the gap with the most people behind it is gap 1 (issue recognition), and the gap that determines whether an automated win actually changes anything is gap 5 (power inside the proceeding). Both are the thin links — and the underlying reason both are thin is the same: the field follows the money, and the money is in gaps 2 and 3, where there are paying customers (small firms, legal aid funders, courts buying procedural tooling), not in gap 1 (no one pays to be told they have a legal problem) or gap 5 (no one pays the unrepresented to be less unrepresented). The honest version of the optimistic story this essay has been telling is that the maturing technology will keep widening the strongest links while doing relatively little for the weakest ones — unless the field deliberately aims at them.

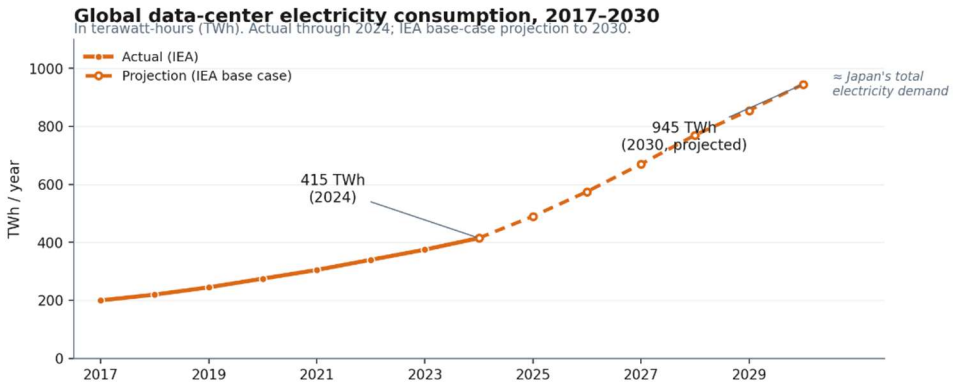
The deeper limitation of the whole framing is that some of what looks like a justice gap is not a service-delivery problem at all. It is a power problem. A tenant facing an eviction with a perfectly designed AI tool and a fully navigated procedure still faces a landlord with counsel, leverage, and the property's deed. Technology helps; it does not equalize. The essay's larger argument — that automation can make justice more inspectable than it has ever been — survives this honestly. But it should be said plainly, especially to readers who would prefer to hear that better tools alone are the answer: more inspectable is not the same as more equal. The first is what the technology delivers. The second has to be insisted on by people, against the institutional gravity that has always preferred the opposite.

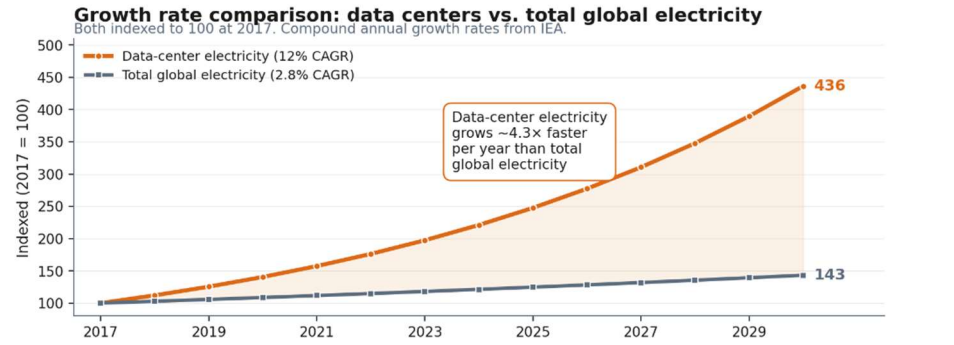
The Externalities the Industry Does Not Like to Talk About

> *Every efficiency gain is paid for somewhere—energy, water, displaced wages, and value shifting from labor to compute owners.*

There is no such thing as a free lunch. Someone always pays. Same as every efficiency gain in legal AI – it's is paid for somewhere. The bills land on energy grids, on water tables, on associate and paralegal wages, and on the share of economic value that flows to human labor versus to compute owners. The optimistic frame — more lawyer time freed for high-value work — is partly true and partly a distraction from harder questions: who captures the surplus, and what do the people freed from the work actually do?

The essay so far has spent almost no time on the costs that do not show up on the legal-services invoice. They are not small, and the case for any of this is weaker if it depends on pretending they are.





Source: International Energy Agency, "Energy and AI," April 2025. IEA reports 415 TWh in 2024, 945 TWh projected by 2030, and a 12% CAGR over the prior five years. Annual values for 2017–2023 reconstructed by back-calculating from the 12% CAGR and the 2024 endpoint; 2024–2030 follows the IEA base-case trajectory between published anchors.

Energy. Global data-center electricity consumption was approximately 415 terawatt-hours in 2024, growing at roughly a twelve-percent compound annual rate since 2017 — more than four times the growth rate of total global electricity consumption. The International Energy Agency's 2025 Energy and AI report projects the figure rising to roughly 945 terawatt-hours by 2030 under its base case, with AI workloads accounting for the largest share of the growth. Electricity consumption from AI-

focused data centers grew approximately fifty percent in 2025 alone. If data centers were a country, they would already be the fifth largest energy consumer in the world. Per-query energy for a frontier-model inference is estimated at 0.3 to 3 watt-hours, roughly three to ten times a Google search of similar intent; a long agentic legal workflow consumes many such queries. The legal-services efficiency gain has a kilowatt-hour denominator, and pretending it does not is the same intellectual dishonesty the essay accused human judges of when it argued their opacity hid their bias.

Water. A typical 100-megawatt AI data center consumes roughly 1.5 to 3 million cubic meters of water annually for evaporative cooling. Hyperscaler water consumption rose twenty-five to forty percent year-over-year in 2024–2025 disclosures. The water is drawn from local sources where the data center sits; the legal-services gain accrues globally; the water cost accrues locally; and the residents of the locality do not get a vote proportional to their loss. This is the textbook shape of an externality, and treating 'AI is more efficient than humans' as a clean claim requires either the externality to be priced or someone to argue, with evidence, that the water and the carbon are worth the legal work they enable.

Wages. The straightforward economic effect of automating the labor of the bottom tiers of a profession is downward pressure on the wages of the people who used to do that labor. The optimistic counter — that lawyers freed from gathering will do more analysis and earn more — is partly true at the top of the profession and largely false at the bottom and middle, where document review, paralegal work, and junior associate hours are the work being eliminated. The wage compression has not yet shown up cleanly in industry-wide data because the senior tier's rates rose enough through 2025 to mask the bottom-tier weakness, but the structural force is there. The 2026 State of the U.S. Legal Market report notes that fees per lawyer are now growing faster than profits — a sign that pricing has gotten ahead of demonstrable value, which is the kind of dynamic that eventually corrects sharply.

The shift of value from labor to compute. Historically, the value created by a legal matter flowed almost entirely to labor: partners, associates, paralegals, support staff, and the firm's equity. This is shifting. Going forward, an increasing share will flow to the owners of the compute that produces the work — through the API providers (Anthropic, OpenAI, Google), through them to chip designers and foundries (Nvidia, AMD, TSMC), and through them to the energy producers and capital markets that finance the data-center build-out. The legal industry's old quarrel was about how the partner-associate share of a matter's value was allocated. The new quarrel will be about how much of the matter's value the legal industry retains at all, versus how much accrues to a stack of suppliers it does not control. This is the substrate

underneath the open-weights argument: a firm that runs its own models keeps more of the value than a firm that rents them.

What people do with freed-up time. The optimistic answer is that lawyers freed from gathering work on harder problems, advise more clients, and live richer lives. Some of that is true. Some of it is also true that the historical pattern, when labor-saving technology arrives in a profession that bills by the hour, is that hours billed stay constant and the work gets denser, not that the worker goes home earlier. A lawyer who was billing 2,000 hours of gathering and analysis in 2024 is, in 2026, billing 2,000 hours of more matters less deeply, with the saved time absorbed by the firm rather than the human. The 'human freed for higher-value work' frame is true in some firms and inverted in others, and which version takes hold is a culture and management question, not a technology question. The same goes downstream of the bill: a client whose legal spend falls by thirty percent does not automatically pass the savings on to anyone. The surplus accrues to whoever has the bargaining power to capture it.

The consumption side. Even if all the technology arrives, the value shifts, the surplus is captured somewhere, and the freed time is taken back by employers — what happens to people's own consumption patterns and to democratic life? A society where the marginal cost of generating fluent expert advice approaches zero is also a society where the things that used to anchor a person's working day — the slow work, the gradual mastery, the relationship with a colleague who taught you something hard — are made optional. Whether that is liberating or hollowing depends, again, on choices outside the technology: how the public domain of legal materials is preserved, how mentorship is funded, how the long tail of human-to-human practice is protected from being eroded by the convenience of asking a model. A democracy whose citizens have abundant, cheap, accurate legal advice is a different democracy than one whose citizens fate depends on unevenly rationed expertise; but a democracy in which all that advice flows through three vendor APIs is also different from one in which the substrate is open. The choice between those last two is one the open-weights section was really about. (Buy that hardware if you can and learn to work with open-weight models.)

None of this argues against the project of transparent, automated legal services described in this essay; it argues against pretending that project is free. The essay has been making the case that an automated legal system, built right, could deliver more justice than the one it replaces. That claim is sustainable only if 'built right' is true, and built right has to include accounting for the externalities the optimistic story does not pay for. The conditions in the prior sections — transparency, audit, recourse, the technical floor, the apprenticeship rebuild, the open-weights option — are the legal-political response. The energy, water, wage, and value-capture costs are

the economic and ecological response, and they have to be addressed in parallel, by different institutions, with the same seriousness.

Where the People Go

> *People remain the accountable node and the judges of novelty—and their deepest new role moves from the bench to the architecture.*

People remain the locus of accountability, the source of judgment in genuine novelty, the human-to-human work of advocacy and counsel, and the choosers of the ends the apparatus serves.

Strip away the timelines and the irreducibly human functions come into focus, and they are not the ones we usually celebrate. They are not drafting, research, or even most of what we call analysis—machines will do those as well or better. They are narrower, deeper, and more about being a person and a citizen than about processing information.

People remain the locus of accountability—the answerable node where consequences land and remedy attaches, because a society that runs on law requires that someone be answerable, and that someone must be capable of bearing the result. People remain the source of judgment in genuine novelty, the cases with no precedent and no pattern, where someone has to choose which value wins. People do the advocacy and counsel that are fundamentally about other people (or people-programmed automations)—persuading a jury, negotiating across a table, sitting with someone through the worst event of their life, fighting for them and bearing the weight with them. And people initially set the goals: the values the whole apparatus serves are initially chosen by humans.

However, the deepest relocation is the new one this essay has been building toward. In a world where the machinery of justice can be made transparent, the most important human work is no longer making each individual call. It is designing the systems to be inspectable, auditing them to be fair, governing the values they encode, preserving real recourse for the people they judge, and ensuring that the new transparency converts into actual accountability and remedy rather than into a more sophisticated way of blaming no one. The human role moves upstream, from the bench to the architecture—and it becomes, if anything, more consequential, because a single design choice now shapes millions of outcomes.

Aiming the Surplus

> *When justice gets cheap, the one decision no machine makes is whether to aim the surplus at the underserved or only at those who could already pay.*

If legal labor and judgment become abundant and cheap, the binding constraint on justice changes, and what we choose to do with the freed-up capacity is a consequential decision in this story. The pull of the market is not a future risk to guard against; it has already decided the order of events. We automated the profitable work first—the corporate NDA that turns around in four minutes—while the eviction tenant, whose answer the same technology could draft just as fast, still faces the court alone. That ordering tracked profit, not need, and it was not a technical necessity but a decision—one we can still reverse. The harder choice, still available, is to close the justice gap on purpose.

Which leads to the part that matters most, and perhaps is the part the industry discusses least, because it is not about law firms or even courts. If legal labor and legal judgment become abundant and cheap, the binding constraint on justice changes, and what we choose to do with the freed-up capacity is the most consequential decision in this entire story. It is human, value-laden, and it could go badly.

The defining scandal of the current legal system is not that lawyers are slow; it is that most people who need legal help never get it. The overwhelming majority of eviction defendants, debtors, immigrants, and parents in custody disputes face the system alone, because lawyers cost more than they can pay. For the first time in history, competent legal help—and competent, consistent, auditable adjudication—could become cheap enough to reach all of them. This may be the single largest opportunity for human good in the entire legal landscape transformation. But it will not happen on its own. The natural pull of the market is to aim all this capability at the clients who could already afford lawyers—large corporations, who get superpowered—while ordinary people are handed a chatbot and called served. The same fork runs through the courts: transparent, auditable, contestable machine justice that finally makes bias visible and correctable, or opaque proprietary systems that automate the old injustice and shield it behind a claim of objectivity. Both forks are technically available. Which one we build is not a technical question, but moral and political.

There is real work to be done, then, and it is not the work of defending the old roles. It is closing the justice gap on purpose, by building the inexpensive, competent help for the people who have never had any rather than only for those who could always pay. It is simplifying the law itself, because once the cost of navigating law collapses, the law's own complexity becomes the new bottleneck, and badly written rules execute badly at scale. It is building the auditing and accountability infrastructure that turns transparency into remedy. It is solving the apprenticeship problem so the profession does not run out of the judgment it cannot automate. And it is guarding the rule of law as a human institution—keeping it legible, legitimate, and answerable to the people it governs, so that automation makes justice more accessible rather

than more alien. This author's tentative view is that, if it is built well, AI may compress the gap between lower and middle-class Americans in their access to expertise — to legal advice, to medical advice, to financial advice that used to be the province of those who could afford it. People who were previously gated out of professional counsel could come closer to the same playing field as those who were not. AI may also compress the wealth gap between lower, middle, and some of the upper classes, putting them on a more even playing field. Whether that economic equalizing produces more mutual care or merely more equally distributed disappointment is an open question that depends less on the technology than on what the technology is asked to do for whom.

The Quiet Room, Revisited

> *The right question was never "will this replace me" but "what is worth keeping, and what will we do with what it frees."*

The unease at that demonstration was real but aimed at the wrong thing. The fear was 'will this replace me.' The better questions were 'what is the me that is worth keeping,' and 'what will we do with everything this frees up.'

Return to that demonstration described at the outset—the agent that did in four minutes what used to take a team a week, and the unease it reportedly left in the room. That unease was real but aimed at the wrong thing. The fear the account captured was *will this replace me*. The better questions were *what is the me that is worth keeping*, and *what will we do with everything this frees up*.

The forecast is this. In two years, the work changes inside the roles: lawyers and judges become editors, supervisors, and owners of machine-generated work, and the economics of the billable hour begin to crack. In ten years, the structure changes: firms thin to a small human shell of accountable principals, engineers, and advocates wrapped around tireless agents, and courts reserve human judges for appeals, novelty, and mercy while machines handle the rest. Throughout, the people who remain are concentrated into the functions that were always the real point of law—being answerable, judging where there is no precedent, advocating as one human for another, and choosing the ends the apparatus serves.

An important truth runs against the reflex to defend the human status quo. The opacity we have prized in human judgment has been hiding its bias, and a transparent machine—built right—could deliver a justice that is more consistent, more inspectable, and more correctable than anything a covertly biased human system has managed. That is not an argument for surrendering justice to machines. It is an argument for taking responsibility for how they are built, because "built right" is doing enormous work in that sentence, and every condition it names is a human choice. We are about to be able to deliver justice that reaches nearly everyone who

has ever needed it and gone without, and to see and fix unfairness that has been invisible for as long as there have been courts. Will we build that version, or merely a faster and better-documented version of the old injustice? That is where the people go—and perhaps it is the most important place they could possibly be.

Appendix: What It Would Take — The Money Case and the Readiness Maps

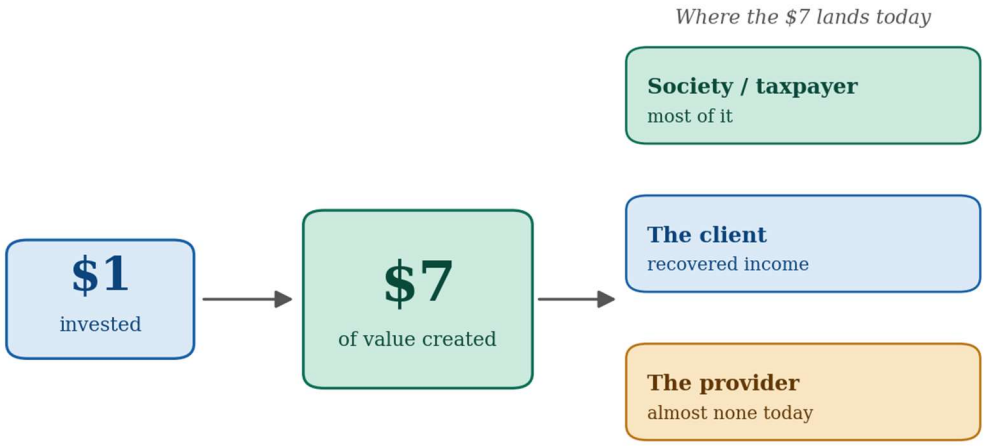
> *Read together, these maps show a pattern: the processes that got automated first are the ones with a payer, not the ones with the most need.*

This appendix does two things. First, it makes the business case that serving access to justice can pay — not only feel good — because "do good" rarely moves capital on its own. Second, it diagnoses, process by process, exactly what it would take for a piece of legal work to run without a human in the loop — and shows that for most high-need civil matters the binding constraint is a policy choice we can change, not a technical limit we must wait out.

Part A — The Money Case

A systematic review of fifty-six economic-impact studies across thirty-nine states (2003–2023) found that every one showed a positive return, averaging about seven dollars of value for every dollar invested in civil legal aid. That is unusually strong evidence — not a cherry-picked study but the entire literature pointing one way. The catch, and the reason A2J has stayed framed as charity, is that most of that value is money saved for society, not money made by whoever does the work. So it helps to separate three distinct money arguments, because they persuade different audiences and only some of them pay a provider's bills.

Follow the $7: civil legal aid returns ~$7 of value per $1 invested



The value is real and proven. The reason it does not pay providers' bills is not that it isn't there — it is that we have not built the mechanisms to capture it. Automation, by dropping cost-per-matter, is what finally lets volume-at-low-price pencil out.

Figure M1. *Follow the $7. Civil legal aid returns roughly $7 of value per $1 invested, but almost none of it is captured by the provider today. The opportunity is not that the value is missing — it is that the capture mechanisms are. Source: Legal Services Corporation, "The Economic Case for Civil Legal Aid" (2024), systematic review of 56 studies.*

The money saved is the best-evidenced argument and the one aimed at governments and courts: preventing an eviction or foreclosure avoids emergency-shelter, medical, and enforcement costs, and clearing a criminal record raises employment and cuts reincarceration — the single most expensive line item in criminal justice. The state-level returns cluster around the $7 average, and the spread itself signals these are real measurements rather than boosterism.

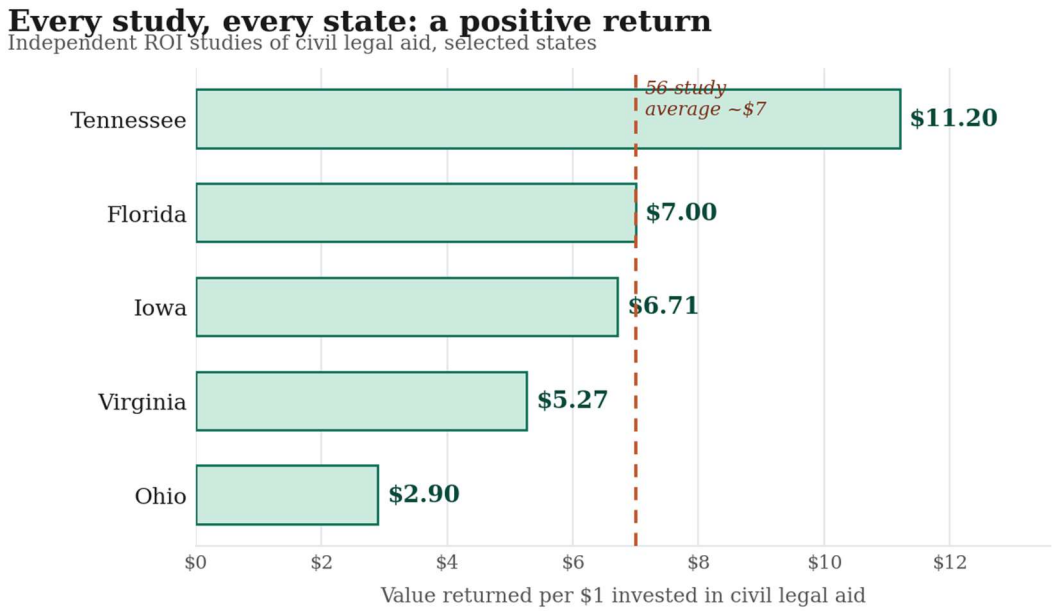


Figure M2. *Return on investment in civil legal aid, selected states. Every independent study reviewed found a positive return. Figures: Tennessee $11.20, Florida ~$7.00, Iowa $6.71, Virginia $5.27, Ohio $2.90; 56-study average ~$7. Sources: state bar foundation and LSC economic-impact studies (2010–2023).*

The honest framing of the for-profit piece — the part that actually pays a founder's bills — is not "charge poor people." It is that automation collapses cost-per-matter far enough that a provider can profit on volume at prices that are attainable for ordinary people, the way an online tribunal resolves a dispute for a few dollars. The latent demand is enormous: conviction-related lost earnings alone run into the hundreds of billions a year, value that re-enters the economy — and a payable services market — once these problems are solved at scale. The three arguments, and who each one convinces, line up like this.

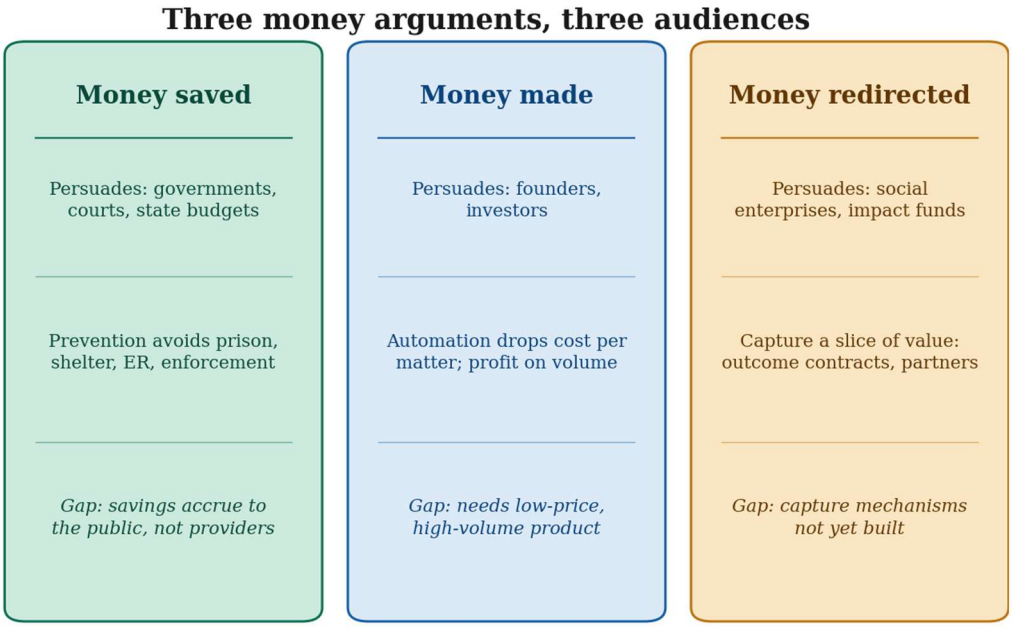


Figure M3. *Three money arguments, three audiences. Money saved persuades governments and courts; money made persuades founders and investors; money redirected persuades social enterprises and impact funds. Each has a different binding gap. Note: ROI studies measure nonprofit legal aid and use varied methods; the for-profit case is a reasoned model for a gap the data implies, not a measured provider return.*

The thesis, then, is not that A2J is secretly lucrative. It is stronger and truer: the value is real and proven; the reason it does not pay providers today is that we have not built the mechanisms to capture it; and automation is exactly what changes that, by dropping cost-per-matter low enough that volume-at-low-price finally pencils out. That reframes A2J from charity-that-saves-money into a market with proven value and a missing business model — which is the order in which we automated legal work, profit before need, and a decision we can still reverse.

Part B — The Readiness Maps

Each map below takes one legal process and asks: for it to run without a human in the loop, what must every step clear? A step passes only if it clears each gate — the data exists and is machine-readable, the governing rule is codified, the output is verifiable, someone has authority to act, the result is accepted, and there is a path to escalate exceptions. Green means a step passes. Amber means it is blocked by a policy choice — which is actionable now. Red means a genuine technical or substantive limit. The payload of each map is the bottom bar: the binding constraint, what it would take to unblock it, and, where it exists, proof the unblocking works. Throughout, “policy” is shorthand for a deliberate legal act — a change to statute, regulation, court rule, or bar rule — as distinct from a technical limit; nearly every

blocker flagged below is exactly such a law or regulatory change, the kind a legislature, agency, court, or bar can make.

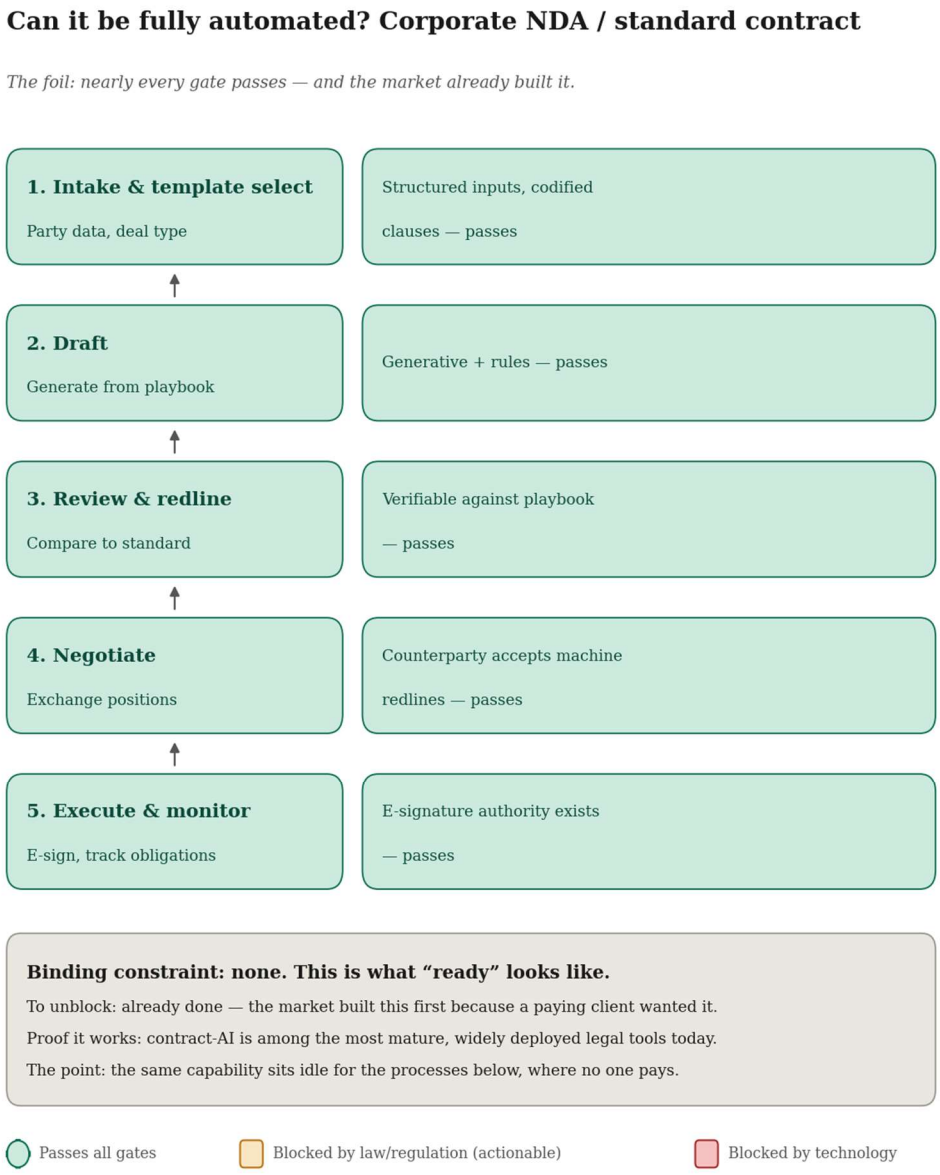


Figure R1. *Corporate NDA / standard contract — the foil. Nearly every gate passes, and the market already built this, because a paying client wanted it.*

Set the next map beside the one above and the argument of this whole essay becomes undeniable. The corporate NDA and the eviction answer have almost identical technical readiness — the same generation, the same verifiability, the same maturity of tooling. The capability that drafts the NDA in four minutes could draft the tenant's answer just as fast. The only variable that differs is revenue: one has a paying client and one does not. So we built the first and left the second undone — not because we could not, but because no one cleared the policy barriers for the process without a

payer. Same readiness, opposite outcomes, and the difference is money, not feasibility.

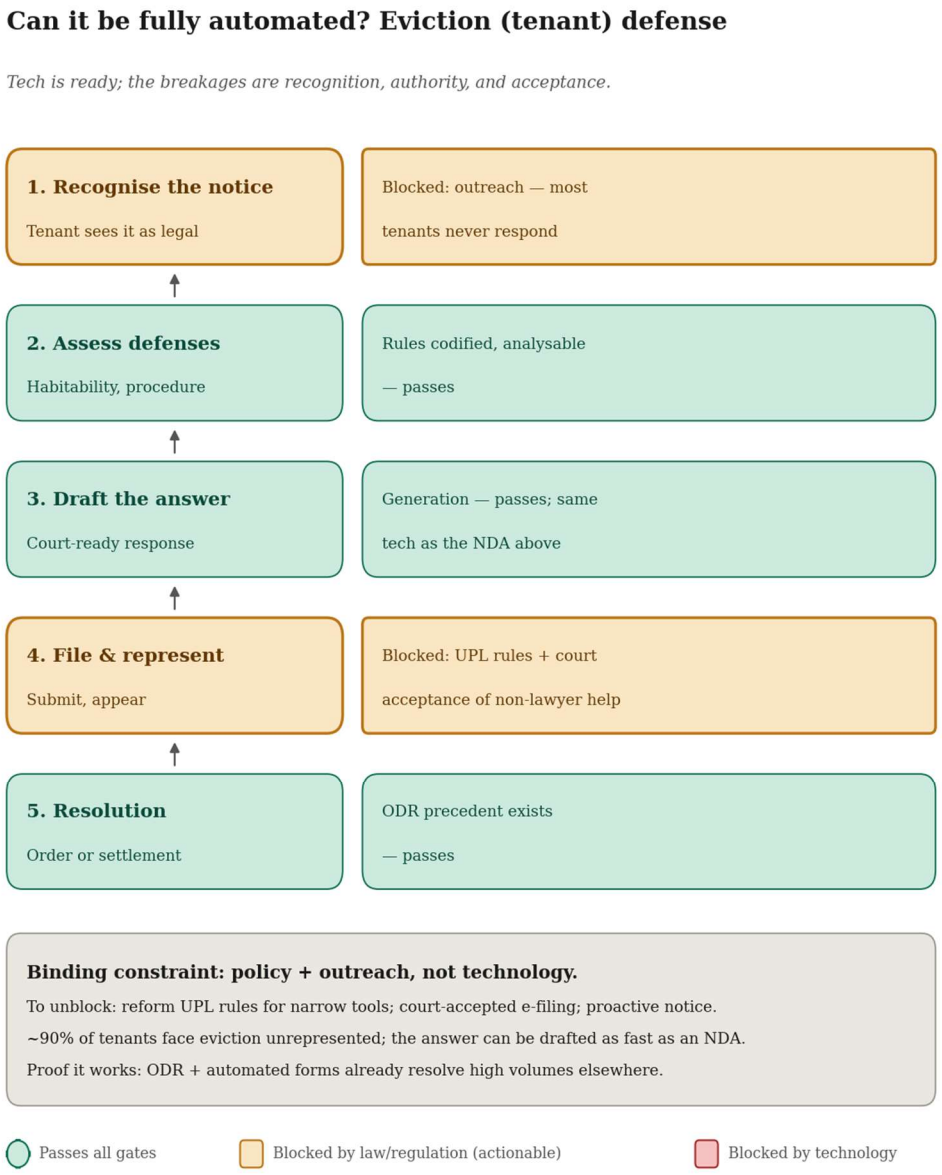


Figure R2. *Eviction (tenant) defense. The drafting gates pass — the same technology as the NDA — but the process is blocked at recognition, unauthorized-practice rules, and court acceptance. Roughly 90% of tenants face eviction unrepresented.*

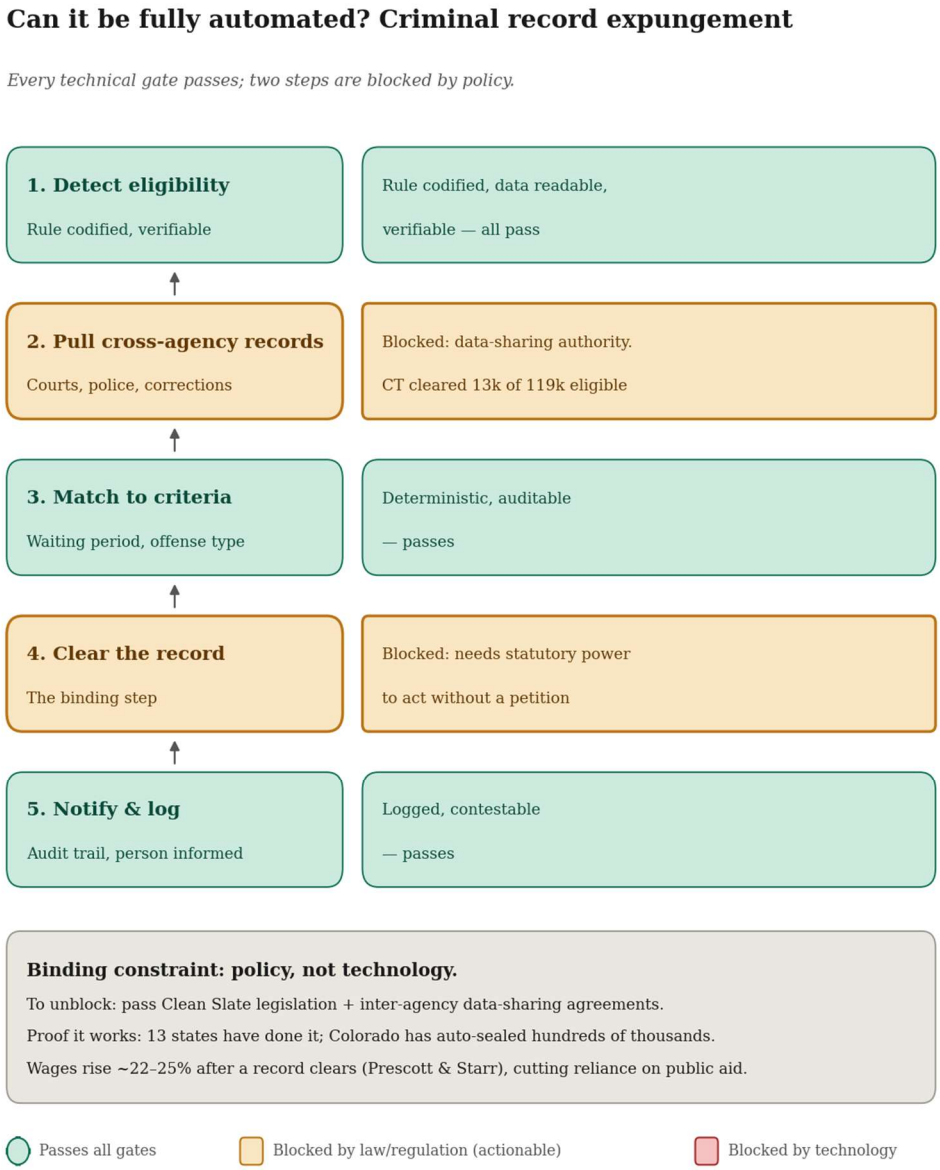


Figure R3. *Criminal record expungement. Every technical gate passes; the blockers are cross-agency data access and statutory authority to clear a record without a petition. Proof: 13 states automate this; Connecticut's backlog (13k of 119k eligible cleared) shows the constraint is data and authority, not capability.*

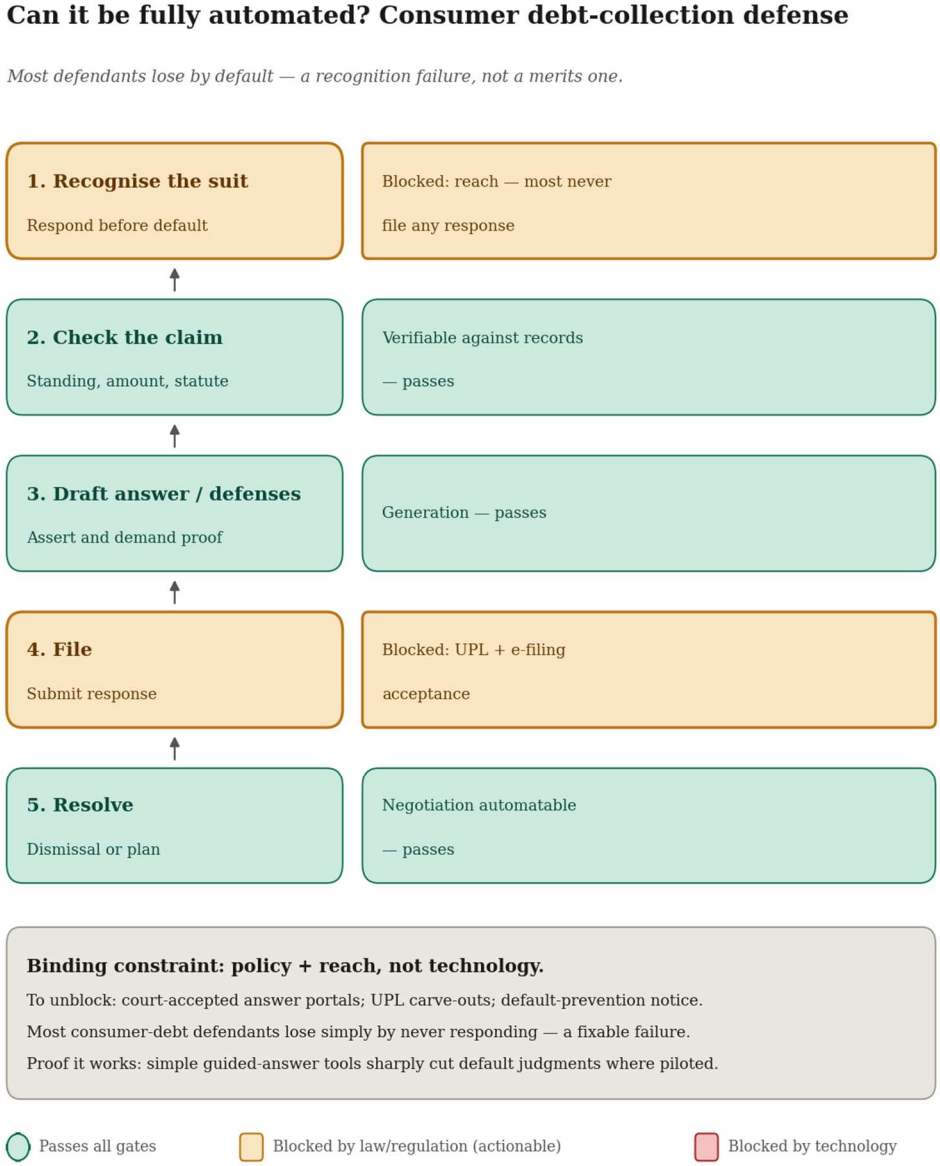


Figure R4. *Consumer debt-collection defense. Most defendants lose by default — a recognition and reach failure, not a merits one. The drafting and resolution gates pass; the blockers are outreach, unauthorized-practice rules, and e-filing acceptance.*

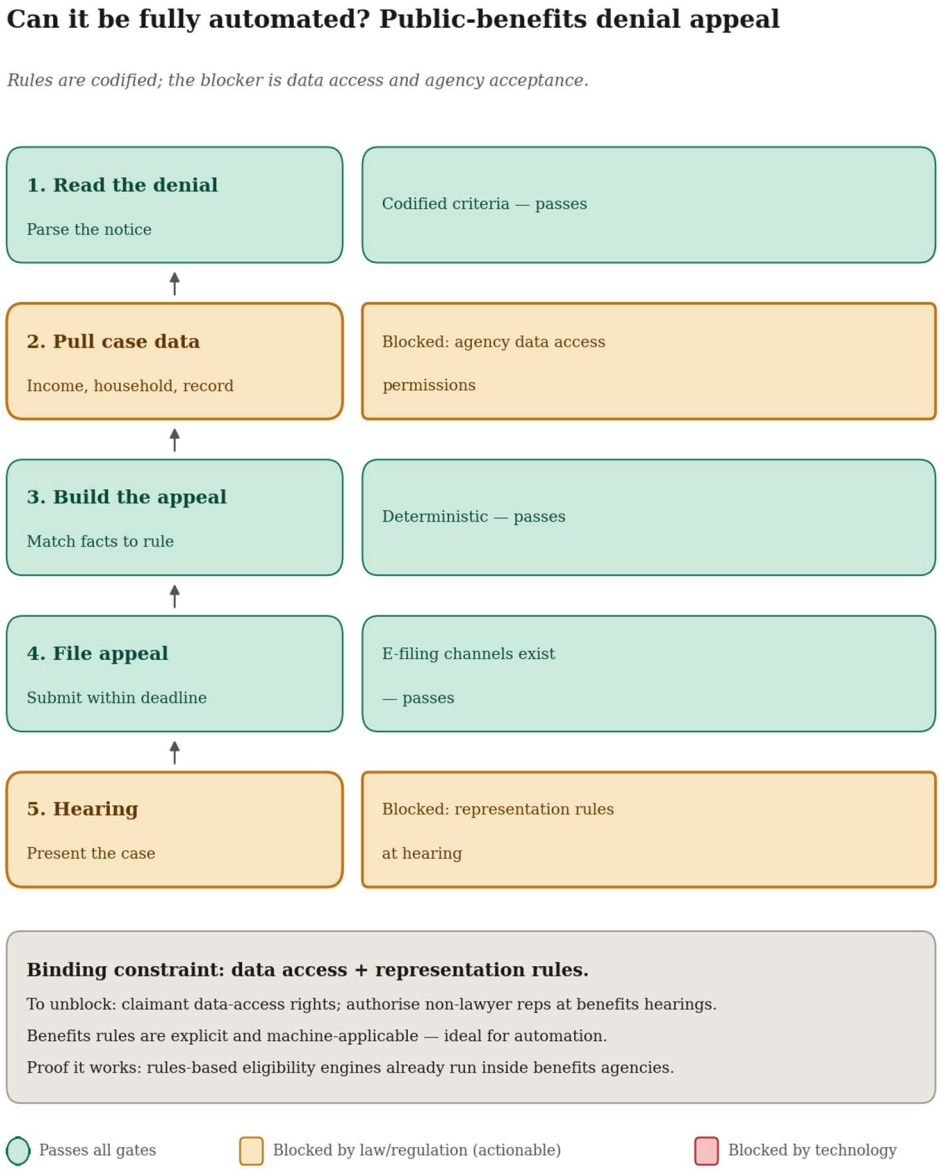


Figure R5. *Public-benefits denial appeal. Benefits rules are explicit and machine-applicable; the binding constraints are claimant data access and representation rules at the hearing.*

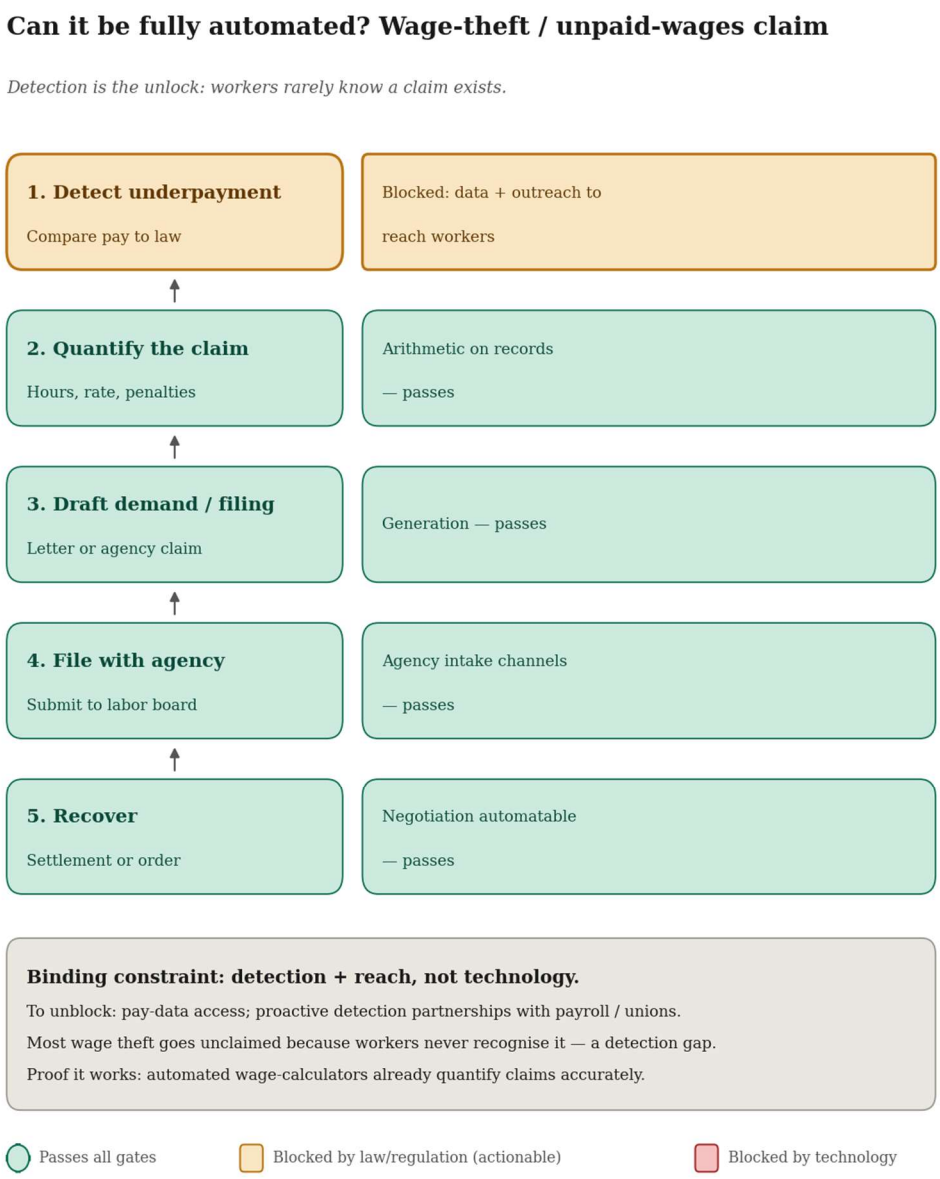


Figure R6. *Wage-theft / unpaid-wages claim. Quantifying and filing automate cleanly; the unlock is detection and reach, since most wage theft goes unclaimed because workers never recognize it.*

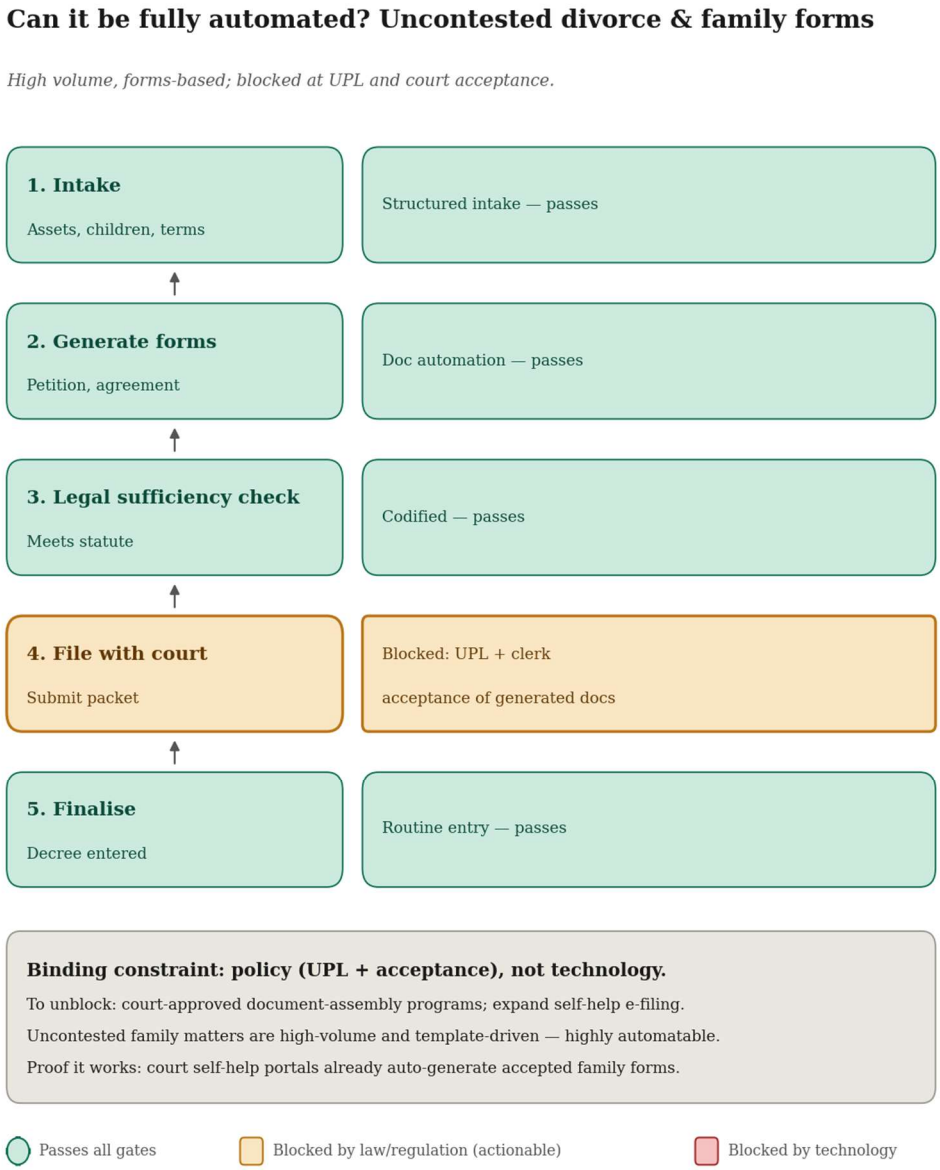


Figure R7. *Uncontested divorce and family forms. High-volume and template-driven; blocked at unauthorized-practice rules and clerk acceptance of generated documents. Court self-help portals already auto-generate accepted forms.*

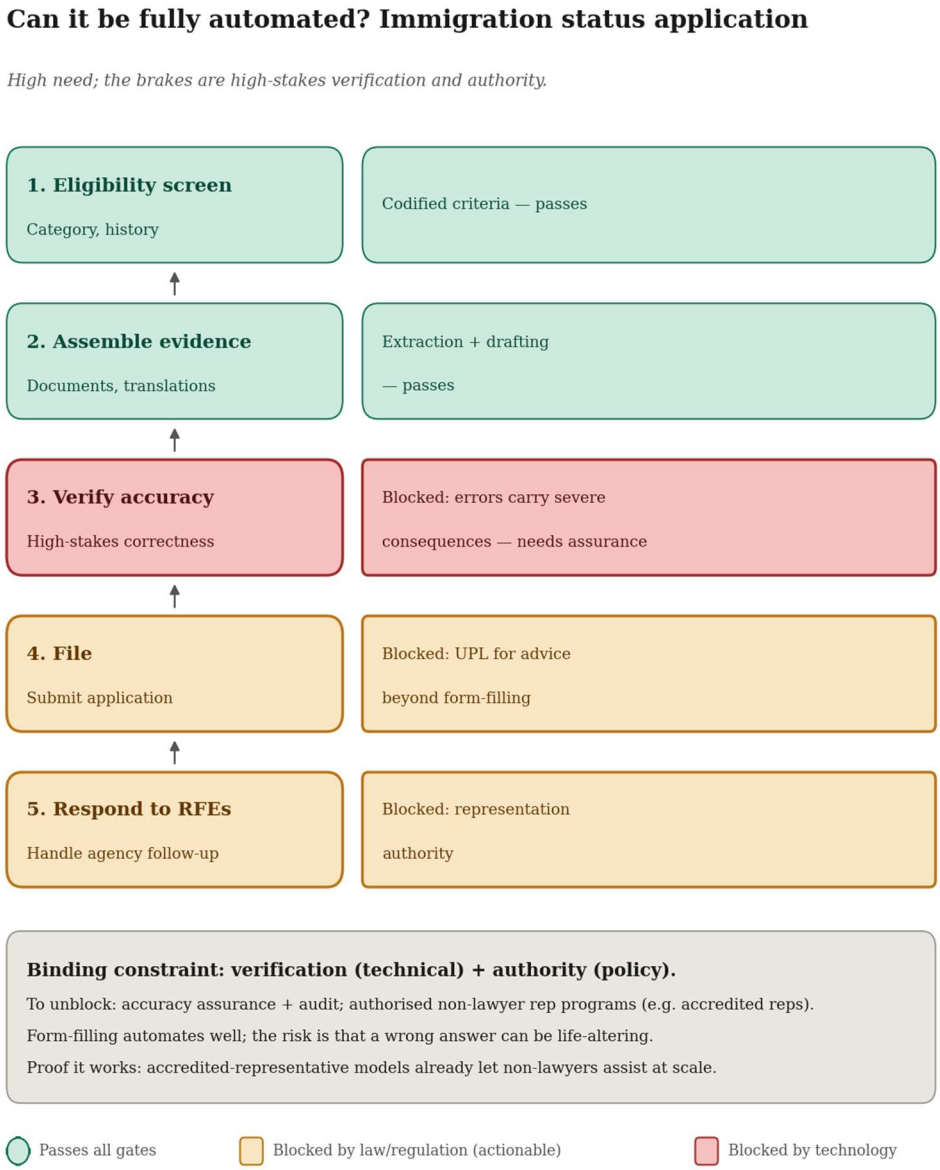


Figure R8. *Immigration status application. Form-filling automates well, but a wrong answer can be life-altering, so verification is a genuine technical brake; authority to advise beyond form-filling is the policy brake. Accredited-representative programs already let non-lawyers assist at scale.*

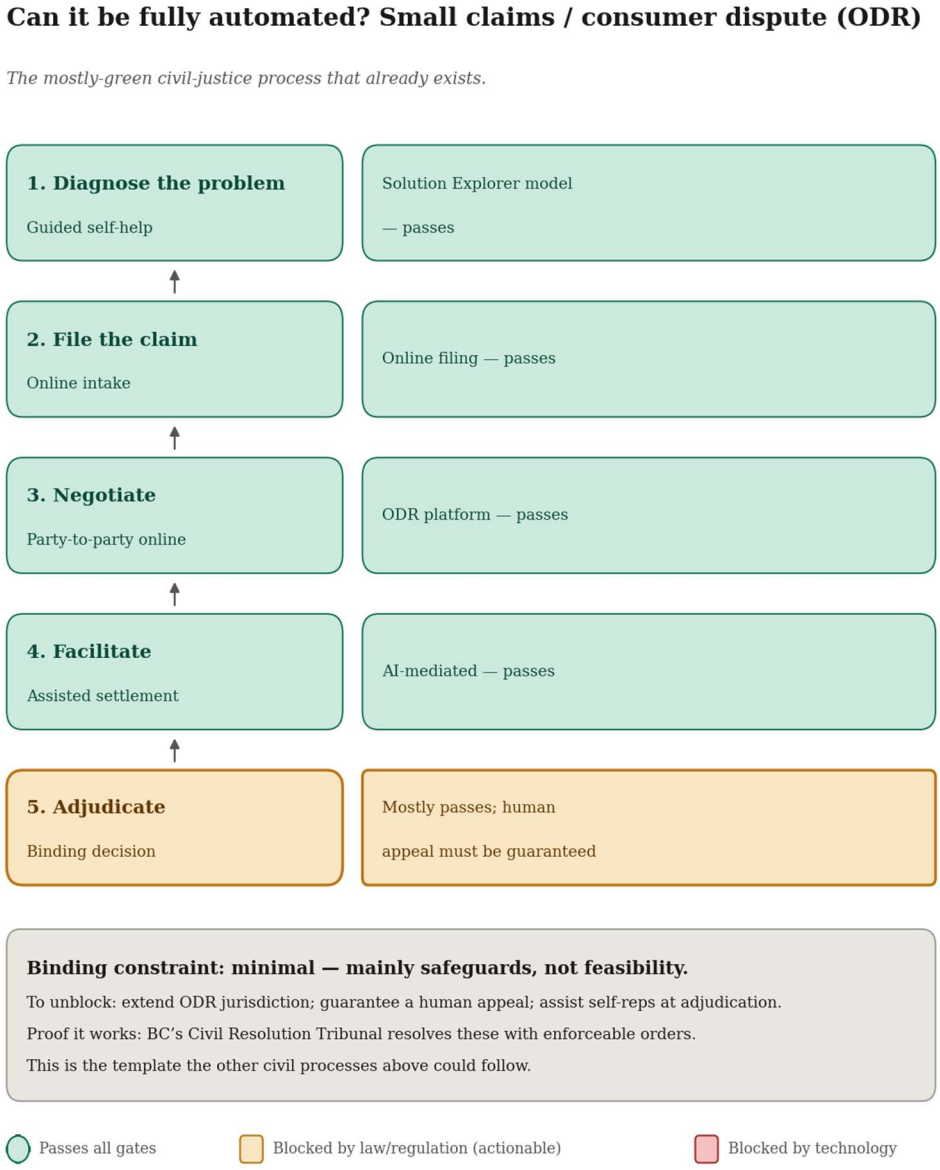


Figure R9. *Small claims / consumer dispute (ODR). The mostly-green civil-justice process that already exists: British Columbia's Civil Resolution Tribunal resolves these end-to-end with enforceable orders. The remaining constraints are safeguards — chiefly a guaranteed human appeal — not feasibility.*

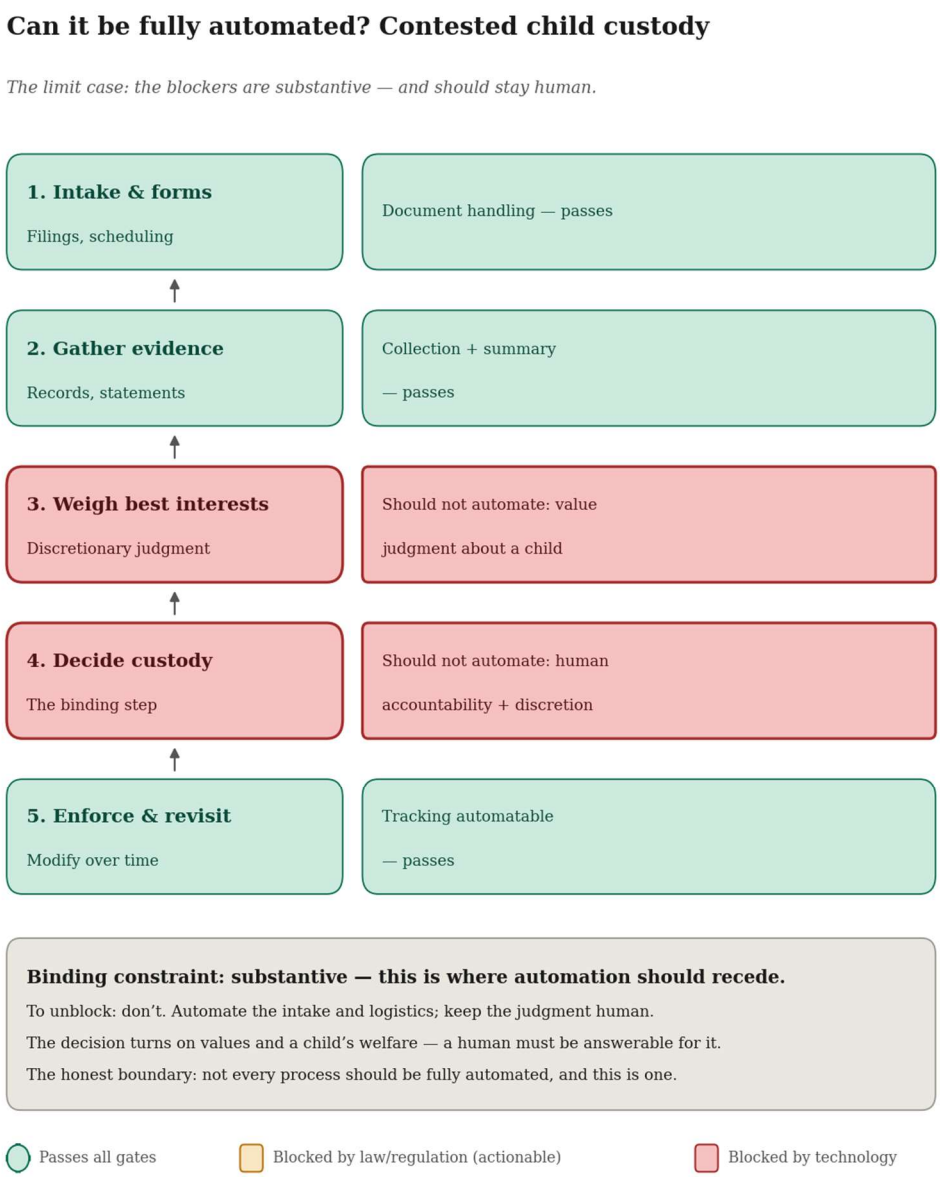


Figure R10. *Contested child custody — the limit case. Intake and logistics automate, but weighing a child's best interests and deciding custody should not: the decision turns on values and a human must be answerable for it. The honest boundary — not every process should be fully automated, and this is one.*

Read as a set, the maps deliver one message. With the lone, deliberate exception of contested custody, almost none of these processes is blocked by technology. They are blocked by data-sharing rules, unauthorized-practice rules, court-acceptance rules, and the absence of outreach — every one of which is a policy choice a legislature, a court system, or a bar association can change. The technical capability that the market has already aimed at the corporate NDA is sitting idle, waiting to be pointed at the eviction tenant, the debtor, the wage-theft victim. Pointing it there is not an engineering problem. It is a decision — the same decision the body of this essay said no machine will make for us.

Notes & Sources

A note on citations: this essay uses inline parentheticals to flag specific empirical claims where the source is doing heavy work, and consolidates the full bibliographic detail here. For readers tracking a particular claim, the corresponding entry below identifies the study, dataset, or institution by short title and provides a URL or DOI where one exists. Where a claim rests on a body of work rather than a single paper, the entry says so explicitly.

> *Where the essay's empirical claims come from, listed so readers can check them directly.*

The empirical claims in this essay draw on a small number of identifiable sources, listed here so readers can check them directly. On lawyer AI adoption: Wolters Kluwer, "2026 Future Ready Lawyer Survey," reporting that more than 90 % of 810 lawyers surveyed across the U.S., China, and nine European jurisdictions use at least one AI tool in daily work (wolterskluwer.com/en/know/future-ready-lawyer-2026); 8am, "2026 Legal Industry Report," documenting adoption rising from 31 % to 69 % in a single year across 1,300+ practitioners; and the ACC/Everlaw GenAI Survey, finding corporate legal AI use rising from 23 % to 52 % over the same period.

On hallucinated authorities: Damien Charlotin, "AI Hallucination Cases" database, damiencharlotin.com/hallucinations (1,497 cases logged as of late May 2026; the database is the source most often cited by Bloomberg Law, the LA Times, and Reuters when reporting the trend). On the 80/20 inversion: Robert J. Couture, quoted in "The Impact of Artificial Intelligence on Law Firms' Business Models," Harvard Law School Center on the Legal Profession, 2025 (clp.law.harvard.edu/knowledge-hub/insights). On the LA County pilot: James Queally, "AI pilot program in L.A. County courts will help judges craft rulings in some cases," Los Angeles Times, Feb. 2026; subsequent CalMatters investigation, May 2026, identifying the tool as "Learned Hand" under a $314,000 contract. On the Legalweek vignette: PlatinumIDS, "The Rise of Agentic AI in Legal Technology," March 2026 (blog.platinumids.com/blog/agentic-ai-legal-technology).

On model honesty: Anthropic, release of Claude Opus 4.8 (28 May 2026), framed as the company's "most honest" model and accompanied by the alignment-science research thread "Teaching Claude Why" (alignment.anthropic.com/2026/teaching-claude-why); contemporaneous coverage in Inc. and TechCrunch. On AI cyberdefense: Google, "Introducing Google AI Threat Defense" (Google Cloud Blog, May 2026); "Sec-Gemini v1" (Google Security Blog, April 2026); and Mandiant's M-Trends 2026 special report on AI risk and resilience.

On the human-justice evidence: Danziger, Levav, & Avnaim-Pesso, "Extraneous factors in judicial decisions," PNAS 108(17) 6889 (2011), with the case-scheduling

caveat in Glöckner, “The irrational hungry judge effect revisited,” Judgment and Decision Making 11(6) 601 (2016). On COMPAS: Julia Angwin, Jeff Larson, Surya Mattu, & Lauren Kirchner, “Machine Bias,” ProPublica, 23 May 2016—the journalistic critique whose technical claims have been extensively debated but whose central concern, the use of a proprietary unauditable algorithm in consequential criminal-justice decisions, remains the better-grounded objection. Sentencing-disparity research after controlling for legally relevant factors is anchored in the U.S. Sentencing Commission’s recurring “Demographic Differences in Federal Sentencing” reports: the 2017 report found Black male offenders received sentences 19.1% longer than similarly situated White male offenders (FY2012–16), 20.4% longer once criminal-history violence was included, and 7.9% longer even within the same guideline range; the 2023 report found a 13.4% gap (FY2017–21). The Afrocentric-features findings are Blair, Judd, & Chapleau, “The Influence of Afrocentric Facial Features in Criminal Sentencing,” Psychological Science 15(10) 674 (2004), and Eberhardt, Davies, Purdie-Vaughns, & Johnson, “Looking Deathworthy,” Psychological Science 17(5) 383 (2006), with subsequent replications; Figure 1 is drawn from the 2017 and 2023 reports. The educational-attitude figures cited in the apprenticeship section come from the Thomson Reuters Institute, “2026 Law Student Pulse Survey” (April 2026, n=1,800+).

On thought leadership cited in this essay: Richard Susskind, Tomorrow’s Lawyers: An Introduction to Your Future (Oxford University Press, 1st ed. 2013, 2nd ed. 2017, 3rd ed. 2023), and How to Think About AI (Oxford University Press, 2025); see also Online Courts and the Future of Justice (2019). Mark A. Cohen, Legal Mosaic (legalmosaic.com) and Mark A. Cohen’s Forbes column, including the recurring ’Law’s Delayed Future’ series; Cohen co-founded Clearspire in 2010 as an early structural alternative to the traditional firm model.

On token economics and the financialization of compute: Reuters, ’Exclusive: China works on AI token futures market, sources say, in race with US’ (28 May 2026); TechCrunch, ’Just like gold and oil, we’ll soon be able to trade AI token futures’ (28 May 2026); Yicai Xing, ’AI Token Futures Market: Commoditization of Compute and Derivatives Contract Design,’ arXiv:2603.21690 (March 2026), which proposes the Standard Inference Token unit. The forty-fold inference-price decline from early 2023 to early 2025 is documented in that paper and corroborated by published vendor pricing histories.

On the open-weights frontier as of May 2026: independent benchmark tracking via Artificial Analysis and the public SWE-Bench Verified leaderboard; DeepSeek V4 Pro at 80.6% SWE-Bench Verified under an MIT license; Kimi K2.6 (Moonshot AI), Qwen 3.6 (Alibaba), Llama 4 Scout/Maverick (Meta), Mistral Large 3 (Apache 2.0), Gemma 4 (Apache 2.0 since April 2026), GLM-5 (MIT). Hardware recommendations

are based on Nvidia and AMD enterprise GPU price ranges and Apple Mac Studio M-series specifications current at writing.

On energy and externalities: International Energy Agency, Energy and AI (2025) and the IEA's 'Key Questions on Energy and AI' executive summary (updated April 2026); Brookings, 'Global energy demands within the AI regulatory landscape' (April 2026); per-query energy estimates from de Vries (2023) and Hugging Face inference studies. On water consumption: hyperscaler sustainability disclosures (Microsoft, Google, Meta, 2024–2025) and the ScienceDirect literature on data-center water footprints. On wage and value-capture dynamics: Thomson Reuters, 2026 State of the U.S. Legal Market.

On the justice gap and the failure-chain framework: Legal Services Corporation, The Justice Gap: The Unmet Civil Legal Needs of Low-Income Americans (2022), reporting that 92% of civil legal problems experienced by low-income Americans received inadequate or no legal help; updated state-level data in subsequent LSC studies. The disaggregation into six failure points draws substantially on Rebecca L. Sandefur, Accessing Justice in the Contemporary USA: Findings from the Community Needs and Services Study (American Bar Foundation, August 2014), which established the empirical foundation for treating the civil justice gap as a chain of distinct failures—and in particular for treating issue recognition (gap 1) as the largest and most under-addressed link; and on Sandefur's meta-analytic finding that the principal advantage of representation lies in procedural navigation rather than substantive legal knowledge 909, 2015), which is what makes gaps 2 and 3 the strongest links for AI substitution. The 2022 LSC Justice Gap Study supplies the headline 92% figure. Margaret Hagan's writing at justiceinnovation.law.stanford.edu informs the framing of the ecosystem more broadly.

On named interventions in the six-gap framework. Gap 2 (triage and pathways): LawDroid LawAnswers AI (lawdroid.com; launched nationwide Sept. 2025; >50,000 legal-aid conversations facilitated); Illinois Legal Aid Online (illinoislegalaid.org); Pro Bono Net LawHelp Interactive (probono.net). Gap 3 (procedural navigation): Upsolve (upsolve.org; founded in Harvard Law School's Access to Justice Lab, 2016; ~18,000 families served, >$860M in debt relieved); JustFix.nyc; Hello Divorce; Documate; Afterpattern. Gap 4 (substantive advice and UPL): Upsolve v. James, 604 F. Supp. 3d 97 (S.D.N.Y. 2022) (preliminary injunction granted on First Amendment grounds), vacated and remanded, No. 22-1345 (2d Cir. Sept. 9, 2025) (UPL statutes content-neutral; intermediate scrutiny applies), case dismissed on remand (S.D.N.Y. Mar. 6, 2026), petition for certiorari pending; Utah Office of Legal Services Innovation regulatory sandbox; Arizona Supreme Court alternative-business-structure rule (effective 2021). Gap 5 (power inside proceedings): Courtroom5; emerging AI-augmented pro-bono tools for asylum and eviction defense.

On the open-source Legal Aid Plugin for Anthropic's Claude (released by LawDroid, 20 May 2026, at LegalAidPlugin.org and on GitHub), separately noted as exactly the kind of vendor-independent A2J infrastructure the Open Weights section recommends: free, auditable, deployable by any legal-aid office. On supporting infrastructure: the Legal Services Corporation's Next Frontier: Harnessing Technology to Close the Justice Gap (2026) and the Technology Initiative Grants program ($4.2M / 32 projects / 22 states most recently announced, Dec. 2025); Stanford's Legal Design Lab (Margaret Hagan, Executive Director; justiceinnovation.law.stanford.edu) and the AI & Access to Justice Initiative; Suffolk Legal Innovation and Technology Lab.

On the Damien Charlotin database update: as of late May 2026 the database has logged 1,497 hallucination cases worldwide and continues to grow; readers consulting this essay after publication should expect the figure to be higher.

On Token Economy pricing as of late May 2026: Anthropic API pricing page (anthropic.com/pricing) for Claude Haiku 4.5 ($1/$5 per million input/output tokens), Claude Sonnet 4.6 ($3/$15), Claude Opus 4.7 ($5/$25), and Claude Opus 4.8 (released 28 May 2026, $5/$25 standard, $10/$50 fast mode). OpenAI API pricing for GPT-5.5 at $5/$30. Google Gemini 3 Pro at approximately $2/$12. DeepSeek V4-Pro permanent discount announced 22 May 2026 (the-decoder.com; platform.deepseek.com), $0.435 input / $0.87 output per million tokens, replacing the temporary promotional rate that had been due to expire 31 May 2026 — making DeepSeek's flagship roughly 11x cheaper on input and 28–34x cheaper on output than the proprietary flagships. On the cost-decline curve for GPT-4-level capability from early 2023 to May 2026: published vendor pricing histories collated by cloudzero.com, costgoat.com, and aipricing.guru. The "sixty-fold reduction in three years" headline figure applies specifically to the cheap-tier production capability band; the same figure does not hold for today's flagship tier in absolute dollar terms.