Can Artificial Intelligence Replace Lawyers and Judges?

A Feasibility and Practicality Assessment of Automating a Political and Expensive Legal System

A speculative research paper, grounded in the empirical and legal-scholarly literature

---

Abstract

The proposal is seductive in its symmetry: if the legal system is expensive because lawyers bill by the hour, and political because judges are elected, appointed, and funded by interested parties, then a machine that neither bills nor votes should cure both diseases at once. This paper tests that proposition against the empirical record.

The conclusion is a qualified and asymmetric one. Automation is already delivering measurable gains in legal *work* — research, drafting, triage, translation, document handling — and there is now randomized-trial evidence that current-generation tools improve both quality and speed. Automation of *adjudication*, by contrast, faces three obstacles that are not engineering problems and will not yield to scaling: (1) formal impossibility results showing that "unbiased" is mathematically underdetermined whenever base rates differ across groups; (2) the demonstrated political valence and fine-tunability of the models themselves, which relocates rather than removes political control; and (3) a legitimacy requirement — that a responsible person give reasons — that is constitutive of adjudication rather than incidental to it.

The most defensible position is that AI will substantially *displace legal labor* while *not replacing judicial authority*, and that the money problem in justice is only partly a labor-cost problem. Where costs have actually fallen for ordinary people, the cause has been process redesign and regulatory reform (online tribunals, unbundled services, relaxed practice monopolies), not automated judgment.

Part I — The Diagnosis: Three Different Diseases

The premise of the question — that the legal system is politically biased and expensive — is broadly supported. But it conflates at least three distinct pathologies, and AI performs very differently against each.

1.1 Money

Money in the selection of judges. In the United States, most judges are state judges, and most state judges face elections. Michael Kang (Northwestern) and Joanna Shepherd (Emory) have produced the most rigorous work on whether that matters. Using a dataset of every state supreme court case in all fifty states over a four-year period, they found that elected judges become more likely to decide in favor of business interests as the amount of campaign contributions received from those interests increases. A later study analyzed 2,345 business-related state supreme court opinions from all fifty states over 2010–2012 alongside more than 175,000 contribution records, finding a strengthening relationship between money and how justices rule in business matters.

Their theoretical account is more interesting than the crude bribery story. They conclude that the intuitive narrative of direct exchanges between individual contributors and judges is too simple; the parties broker the connections, and it is partisan coalitions rather than elections as such that carry money's influence. They also report an asymmetry:

Republican judges respond only to conservative money, while Democratic judges are influenced by both liberal and conservative sources, so the net effect runs in a conservative direction.

The honest caveat. This literature is contested. A replication study using the same 2,345-case dataset found that Shepherd's main findings generally replicate, but that contributions from non-business and anti-business interests likewise increase the probability of votes favoring non-business interests — meaning influence is not the exclusive property of business, which complicates the normative conclusion. A machine that eliminates *business* influence and *labor/plaintiff* influence simultaneously eliminates something we might just as accurately call "responsiveness."

Money as exclusion. The larger money problem is not corruption but absence. The Legal Services Corporation's 2022 Justice Gap Study found that low-income Americans received no or inadequate legal help for 92% of the civil legal problems that substantially affected them, with 74% of low-income households experiencing at least one such problem in the past year and 39% experiencing five or more. This is up from 86% in the 2017 study, which used a slightly different methodology. Rebecca Sandefur's research found that only about 14 percent of civil justice problems are ever taken to a court or hearing body at all, and Stanford researchers report that in roughly three-quarters of civil cases nationwide at least one party is unrepresented — about 15 million cases a year.

This distinction matters enormously for the AI question. A robot judge does nothing about the 86% of problems that never reach a tribunal. An AI *advocate* might.

1.2 Politics

Judicial politicization operates through appointment, election, ideological sorting, and doctrinal drift. The empirical literature on ideological voting in appellate courts is vast and largely uncontested in its central claim: party of appointing president predicts votes in ideologically salient cases. This is the pathology that most people mean when they imagine an impartial machine.

1.3 Arbitrariness (which is not the same thing)

The third pathology is inconsistency — the same case decided differently by different adjudicators, or by the same adjudicator at different times.

The landmark study is *Refugee Roulette*. Analyzing 133,000 decisions by 884 asylum officers over seven years, 140,000 decisions of 225 immigration judges, 126,000 Board of Immigration Appeals decisions, and 4,215 appellate decisions, the authors found significant disparities in grant rates even when different adjudicators in the same office each handled large numbers of applications from nationals of the same country — such that the most important moment in an asylum case is often the instant a clerk randomly assigns it. Grant rates correlated with representation status, the gender of the immigration judge, and the judge's pre-appointment work experience.

The famous "hungry judges" study is the other pillar — and it deserves a warning label. Danziger, Levav, and Avnaim-Pesso reported in *PNAS* that the percentage of favorable parole rulings dropped from roughly 65% to near zero within each decision session and returned abruptly to about 65% after a food break. But Weinshall-Margel and Shapard responded that case ordering is not random: the board tries to complete all cases from one prison before a break, and within each session unrepresented prisoners usually go last and are less likely to be paroled — so the post-break peak is likely an artifact of case ordering rather than evidence that meal breaks affect decisions.

This is directly relevant to our question, and one legal scholar has made the point explicitly. Konstantin Chatziathanasiou's article in the *German Law Journal* is titled, pointedly, "Beware the Lure of Narratives: 'Hungry Judges' Should Not Motivate the Use of 'Artificial Intelligence' in Law," noting that the finding "has entered common

wisdom" and is "broadly and uncritically used as an argument". If the empirical case for replacing judges rests substantially on a finding whose identifying assumption failed, the case is weaker than its advocates believe.

The analytic payoff of this section: money-in-selection, ideological politics, and arbitrariness are three different problems. AI plausibly reduces the third. It does little about the first. And — as Part III argues — it may make the second *worse* in a specific and underappreciated way.

Part II — Technical Feasibility: What the Machines Can Actually Do

2.1 Legal work is not one job

"Replacing lawyers" bundles together tasks with wildly different automation profiles:

Task	Nature	Automation difficulty
Legal research and citation retrieval	Information retrieval over a closed corpus	Low–moderate
Document review, discovery, due diligence	Classification at scale	Low
Drafting (contracts, motions, letters)	Constrained generation	Low–moderate
Legal analysis and strategy	Reasoning under uncertainty	Moderate–high
Fact-finding and credibility assessment	Perception + inference about persons	High
Discretionary judgment (sentencing, equity, remedies)	Value choice	Not a capability problem
Conferring legitimacy on an outcome	Institutional authority	Not a capability problem at all

The last two rows are the crux. They are frequently smuggled into feasibility discussions as if they were engineering milestones. They are not.

2.2 The benchmark evidence, corrected

The most-cited capability claim — GPT-4's "90th percentile" bar exam performance — does not survive scrutiny. Eric Martínez's peer-reviewed reanalysis in *Artificial Intelligence and Law* found that the 90th-percentile figure derived from February administrations of the Illinois Bar Exam, which are heavily skewed toward repeat test-takers who had failed in July and score well below the general population; data from a July administration suggests GPT-4's overall percentile was below the 69th, and roughly the 48th on essays. Against first-time test takers, performance is estimated at roughly the 62nd percentile overall and 42nd on essays. Against those who actually passed — that is, licensed attorneys — it falls to about the 48th percentile overall and the 15th percentile on essays.

The essay collapse is the signal worth attending to. Multiple-choice performance measures recall and pattern-matching; essay performance measures the thing adjudication actually requires. Martínez is careful not to overclaim in the other direction: the improvement from GPT-3.5 to GPT-4 remains impressive even on the corrected numbers.

2.3 Retrieval does not solve fabrication

Legal AI vendors marketed retrieval-augmented generation (RAG) as the cure for hallucination. The first preregistered independent evaluation, by Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher Manning, and Daniel Ho of Stanford's RegLab — published in the *Journal of Empirical Legal Studies* — found otherwise. Against provider claims of "eliminating" or "hallucination-free" citations, the study found that Lexis+ AI, Westlaw AI-Assisted Research, and Ask Practical Law AI each hallucinate between 17% and 33% of the time, though less than general-purpose GPT-4. Performance varied sharply: Lexis+ AI answered 65% of queries accurately; Westlaw's AI-Assisted Research was accurate 42% of the time but hallucinated nearly twice as often as the other legal tools; Ask Practical Law AI gave incomplete answers on more than 60% of queries.

Plain-language version: retrieval means the system looks up real documents before answering. It reduces the invention of nonexistent cases. It does not prevent the subtler and more dangerous failure — citing a real case for a proposition it does not support, or citing a real case that has been overruled.

Technical version: RAG constrains the *retrieval* step but not the *synthesis* step. The generator remains free to produce claims unsupported by, or contradicted by, retrieved context. Groundedness is a property of the mapping from retrieved evidence to asserted proposition, and nothing in the RAG architecture enforces it. The paper's own limitations section notes the closed nature of these systems: Lexis and Thomson Reuters restrict interface access, requiring extensive manual evaluation, and the legal AI product space remains largely closed to independent audit.

2.4 The field trial: the courts themselves

The strongest available evidence on real-world error rates is not a benchmark but the litigation record. Damien Charlotin, a research fellow at HEC Paris, maintains a public database of court decisions in which a tribunal found that a party relied on AI-fabricated material. Trackers built on that database report a striking trajectory: roughly 200 cases in mid-2025, 719 by January 2026, 1,227 by early April 2026, and 1,598 by June 2026 — with the database's inclusion bar limited to decisions where a court explicitly found or clearly implied reliance on hallucinated material, so accusations alone do not count. As of early July 2026, one reading of the database counted 1,668 cases — 1,163 in the US and 59 in the UK, with 653 involving practising lawyers and 975 involving self-represented litigants.

Sanctions have escalated from the original $5,000 fine in *Mata v. Avianca* to $15,000 per attorney in a federal court of appeals and the first indefinite bar suspensions tied to AI filings.

Two cautions. First, these numbers come from vendor-maintained trackers built atop Charlotin's database; the underlying database is the citable source, and the aggregate counts should be treated as approximate and fast-moving. Second, the split between lawyers and pro se litigants is itself informative: the majority of incidents involve unrepresented people, which is precisely the population AI advocates promise to serve.

2.5 The counter-evidence: AI genuinely improves legal work

It would be dishonest to present only the failures. Daniel Schwarcz and colleagues ran the first randomized controlled trial of current-generation legal AI, assigning 127 upper-level law students at Minnesota and Michigan to complete six realistic legal tasks with a RAG-powered tool (Vincent AI), a reasoning model (o1-preview), or no AI. Both tools significantly enhanced legal work quality — a marked contrast with earlier research on GPT-4-era models — while preserving efficiency gains, with statistically significant productivity gains of roughly 50% to 130% in five of six tasks and particularly strong effects on complex tasks like drafting persuasive letters and analyzing complaints. Critically for the hallucination question: o1-preview improved analytical depth but produced some hallucinations, whereas participants using Vincent AI produced roughly the same number of hallucinations as participants using no AI at all.

This is the single most important finding for the *lawyers* half of the question. It suggests that architecture matters, that the technology is improving on a measurable trajectory, and that the realistic near-term outcome is augmentation with large productivity multipliers rather than wholesale substitution.

2.6 Machines as judges: the direct experiments

Two recent studies test the adjudicative question head-on, and they point in opposite directions.

Against replacement. Eric Posner and Shivam Saran (University of Chicago) replicated a 2×2 factorial experiment originally run on 31 U.S. federal judges — a simulated war-crimes appeal varying (a) how sympathetically the defendant is portrayed and (b) whether the lower court followed precedent. They found GPT-4o strongly affected by precedent but not by sympathy — similar to the student subjects, and the opposite of the professional judges, who were influenced by sympathy. In their framing, the LLM behaved like a formalist.

This result is genuinely ambiguous, and its ambiguity is the point. If you think sympathy is a bug, the machine wins. If you think the professional judges were exercising the equitable sensitivity that distinguishes adjudication from rule-execution — which is, roughly, the view of every equity tradition since Aristotle's *epieikeia* — the machine is failing at the job.

For replacement. A 2026 study in *Justice Quarterly* by Gazal-Ayal, Elyoseph, and Solomon compared LLM sentencing decisions to those of 123 retired judges on two fictional cases involving rape and violence, testing GPT, Gemini, and Claude under zero-shot, few-shot, and chain-of-thought prompting. The LLMs showed greater consistency, producing significantly lower sentence disparity than the judges; using the judges' average sentence as a conservative accuracy benchmark, models that are both more consistent and closer to that average could reasonably be considered at least as accurate as individual judges.

Note the epistemological move required here, which the authors flag honestly: they must use the human average as ground truth because the "correct" sentence is unknown. This is the central methodological trap of the whole field. There is no external criterion of correctness for a sentence, so machine "accuracy" is defined by proximity to aggregate human judgment — which means an AI judge is, by construction, a device for regressing everyone toward the historical mean. That is a real improvement over lottery-like variance. It is not the elimination of bias; it is the *averaging* of bias.

2.7 Prediction is not adjudication

A large literature claims high accuracy in "predicting" court decisions. The foundational study by Aletras and colleagues reported 79% accuracy predicting European Court of Human Rights outcomes from case text, with the formal facts of a case emerging as the most important predictive factor.

Masha Medvedeva, Michel Vols, and Martijn Wieling substantially deflated this. They distinguish outcome *identification*, outcome-based judgement *categorisation*, and outcome *forecasting*, and note that although Aletras et al. excluded the verdict section, they used remaining judgment text that often still contained explicit references to the verdict — meaning the task performed was outcome identification, not prediction. Their own reanalysis found that predicting future cases from past cases degrades performance substantially (average accuracy 58–68%), and that roughly 65% average accuracy is achievable using only the surnames of the judges hearing the case.

Plain-language version: much of the "AI predicts court rulings" literature was reading the answer off the back of the exam paper. And to the extent genuine prediction succeeds, it often succeeds by learning *which judge you drew* — which is a measurement of the arbitrariness problem, not a solution to it.

Part III — The Bias Argument, Examined

This is where the popular case for AI judges is weakest, and where the technical literature is most decisive.

3.1 "Unbiased" is mathematically underdetermined

The COMPAS controversy produced the most important formal result in this area. ProPublica argued the recidivism tool was racially biased because Black defendants were disproportionately mislabeled high-risk; Northpointe responded that the tool was calibrated within each racial group, meaning that among defendants at any given score, the proportion who actually recidivated was similar across races.

Both were correct. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, and independently Alexandra Chouldechova, proved that the two positions cannot be reconciled. Chouldechova showed that the differences in false positive and false negative rates cited as evidence of racial bias are a direct mathematical consequence of applying an instrument satisfying predictive parity to a population in which recidivism prevalence differs across groups — the group with higher prevalence will experience a higher false positive rate no matter what, holding positive predictive value constant. The disparities cannot, as a mathematical fact, be avoided. Kleinberg and colleagues constructed a proof that the two notions of fairness are incompatible.

Plain-language version: you can build a system that gives the same score the same meaning for everyone, or you can build one that makes mistakes at equal rates for everyone. You cannot have both when the underlying rates differ between groups. This is not a limitation of current technology. It is a theorem.

Why this is decisive for our question: if "eliminate bias" has no coherent single meaning, then every deployed system silently embeds a *choice* about which fairness criterion to sacrifice. As the literature puts it, these impossibility results imply that any deployed system necessarily embodies particular value judgments about which groups' interests to prioritize. The choice is political. It has simply moved from a courtroom, where it is contestable and visible, into a specification document, where it is neither.

3.2 The training data is the past

An adjudicative model trained on decided cases learns the empirical regularities of prior adjudication, including its discriminations. This is not a data-hygiene problem solvable by filtering protected attributes; correlated proxies (zip code, employment history, prior contacts with the system) reconstruct them. Sonja Starr's work on evidence-based sentencing made this point about statistical risk scores well before the LLM era, and Pasquale's survey notes the accumulated literature on the racial consequences of automatically calculated sentencing risk scores having disparate impacts on minorities.

There is a deeper jurisprudential issue. Common law systems progress partly by *departing* from precedent — *Brown v. Board* is the canonical example. A system optimized to reproduce the distribution of past decisions is structurally incapable of the kind of correction that constitutes legal progress, and would have coded *Plessy* as the correct answer with high confidence.

3.3 Language models have politics — and they are tunable

This is the finding that most directly refutes the premise of the question.

David Rozado, publishing in *PLOS ONE*, administered 11 political orientation tests to 24 open- and closed-source conversational LLMs. A significant majority produced responses diagnosed as left-of-center. Foundational (non-instruction-tuned) models tended to produce mostly incoherent though politically neutral responses. The finding has

replicated widely: the pattern has been found across dozens of models and multiple national contexts, first formalized for GPT-2, and replicated using questionnaires from G7 countries and across US, Brazilian, and UK contexts — robust enough to be considered established. One plausible mechanism is training data: the dominant narratives in a large, Western-skewed digital corpus may skew outputs toward positions aligned with progressivism and social liberalism.

The second half of Rozado's study is the alarming part. Using supervised fine-tuning on politically aligned data, he produced left-leaning, right-leaning, and deliberately depolarizing variants of the same base model, and demonstrated that fine-tuning with politically aligned data shifts ideological outputs, highlighting the malleability of LLM biases.

The consequence: the political orientation of an AI judiciary would be a *tuning parameter*. Whoever controls the fine-tuning corpus, the reward model, and the system prompt controls the ideological center of gravity of the entire judiciary — invisibly, at scale, and without the friction of confirmation hearings, judicial elections, or the need to persuade a single colleague.

Consider what this does to the money problem. Today, influencing state judicial outcomes requires spending across many races over many cycles, with contested and observable results. Under an automated judiciary, influencing outcomes requires capturing one procurement decision or one model-alignment process. The attack surface does not disappear. It *concentrates* — which is the opposite of what the proposal promises.

3.4 Opacity and the trade-secret problem

State v. Loomis (2016 WI 68) is the cautionary precedent. Eric Loomis was sentenced partly in reliance on a COMPAS risk assessment. Because the methodology behind COMPAS is a trade secret, only the risk estimates are reported to the court; Loomis argued that reliance on it violated due process by preventing him from challenging its scientific validity, and separately that it unconstitutionally considered gender. The Wisconsin Supreme Court held 7-0, with two concurrences, that if used properly and subject to specified limitations and cautions, consideration of a COMPAS risk assessment at sentencing does not violate due process. The developer had not designed the tool for sentencing use and would not disclose the algorithm. The U.S. Supreme Court denied certiorari in June 2017, declining the opportunity to rule on whether it violates due process to sentence someone based on a risk-assessment instrument whose workings are protected as a trade secret.

Loomis establishes the pattern the automation debate must confront: courts have proven willing to accept opaque commercial systems on the reasoning that a human retains formal authority — while the practical question of whether that human meaningfully departs from the score goes unexamined.

3.5 Automation bias: the human-in-the-loop may be decorative

Every serious regulatory framework retains a human decision-maker. Whether that retention is substantive is an empirical question, and the answer from adjacent domains (aviation, medicine, credit) is that humans under time pressure defer to system outputs, particularly when departing from them requires written justification and creates reversal risk. A judiciary in which AI recommends and judges approve at a 97% rate is functionally an AI judiciary with a human liability shield. Notably, the reverse workflow is already documented in China: in Shenzhen, judges make an initial decision, an LLM generates supporting reasoning, and the judges then revise it into a final judgment — suggesting a role in articulating reasoning rather than determining outcomes. Which of these two workflows becomes standard is arguably the single most consequential design question in the field.

Part IV — Would AI Actually Remove Money from Justice?

4.1 Where the money actually is

The cost of a legal system is not principally the cost of judges. Judicial salaries are a small fraction of total system cost. The dominant costs are (a) lawyer time on both sides, (b) discovery, (c) delay and its downstream economic consequences, and (d) the opportunity costs borne by parties. Automating the *judge* touches (c) and a little of (d). Automating *lawyers* touches (a) and (b) — which is where the money is.

This is why the framing of the original question is slightly inverted. Automating advocacy would do far more for affordability than automating adjudication, and it is also the part that is technically closer and legally easier.

4.2 The asymmetric adoption problem

Marc Galanter's classic account of why "the haves come out ahead" turns on repeat players' structural advantages: they can play for rules rather than outcomes, absorb losses, and invest in specialized capability. AI is, in the near term, an amplifier of exactly this asymmetry. Frontier capability costs money; the best deployments require data infrastructure, evaluation pipelines, and legal engineering staff. A corporate defendant with a tuned litigation stack facing a pro se plaintiff with a free chatbot is not a leveled playing field — and the hallucination data above suggest that pro se litigants are already bearing the sharp end of the error distribution, since 975 of the tracked cases involved self-represented litigants against 653 involving practising lawyers.

Whether AI narrows or widens the justice gap therefore depends almost entirely on deployment policy, not on capability. Capability is available to everyone; *reliable* capability is not.

4.3 What has actually reduced cost

Here the evidence is encouraging, and instructive, because the wins have not come from automated judgment.

Online tribunals. British Columbia's Civil Resolution Tribunal, the first tribunal in Canada to implement online dispute resolution and often described as the first online tribunal integrated into a public justice system, reports that 94 per cent of disputes settle before reaching a tribunal member, with the process redesigned around a twelve-year-old's rather than a lawyer's reading age, and only 6% of claims reaching adjudication. A user survey found that for many participants the CRT provided an accessible, convenient, and proportional means of resolving legal problems where they may have had no other realistic alternative, and a majority of those with prior experience of traditional courts found it an improvement — though some found the process confusing to navigate. Richard Susskind, who chaired the UK Civil Justice Council's ODR Advisory Group, has argued that the great social benefit of ODR lies in resolving disputes parties would otherwise never bring to court at all.

Note what did the work here: guided pathways, staged negotiation, plain-language design, and low-cost human adjudication at the end. Not machine judgment.

Regulatory reform. In 2020, Arizona abolished the rule barring non-lawyer ownership of law firms, and Utah launched a regulatory sandbox permitting entities to test service models that would otherwise violate unauthorized-practice-of-law or ethics rules. Early data from both states show minimal consumer harm and growing innovation, though critical questions remain. Both reforms were driven by a documented access-to-justice crisis exacerbated by rules that stymie innovation and prop up a lawyers' monopoly over legal services. Utah's sandbox is time-limited, running until August 2027.

The lesson. Cost reduction in justice has come from redesigning *process* and relaxing *market* restrictions. AI is a powerful accelerant for both. It is not, on the evidence, a substitute for either.

Part V — Legal and Constitutional Feasibility

Even if a machine could adjudicate well, the law of most jurisdictions currently forbids it — and, importantly, the forbidding is getting *more* explicit as capability grows, not less.

5.1 European Union

The AI Act (Regulation 2024/1689) classifies as high-risk AI systems intended to be used by a judicial authority or on their behalf to assist in researching and interpreting facts and the law and in applying the law to a concrete set of facts, or to be used in a similar way in alternative dispute resolution (Annex III, point 8(a)). Recital 61 explains the rationale and states the limit in terms that leave little room: such systems are high-risk given their potentially significant impact on democracy, the rule of law, individual freedoms, and the right to an effective remedy and a fair trial; AI tools can support the decision-making power of judges or judicial independence but should not replace it — the final decision-making must remain a human-driven activity. Purely ancillary administrative uses — anonymisation or pseudonymisation of decisions, internal communication, administrative tasks — are carved out.

There is an important loophole scholars have flagged: the majority of judicial decision-support systems may be exempted from high-risk classification under the Article 6(3) derogation for systems that do not materially influence the outcome of decision-making, including those performing a narrow procedural task. Whether a "narrow procedural task" that shapes 90% of outcomes is materially influential is the litigation waiting to happen.

The European Convention on Human Rights supplies the deeper constraint. Article 6 guarantees a fair trial before an independent and impartial tribunal, with a reasoned decision. The scholarly consensus is that since the justice system is a social process, an AI assistant is preferable to a situation in which a robot judge completely replaces human judges, and AI assistants can better serve legitimacy and guarantee a fair trial.

5.2 France: the counter-move

France went further than restraint — it criminalized a *precondition* of predictive justice. Article 33 of the 2019 Justice Reform Act provides that the identity data of judges and members of the judicial registry cannot be used to aid in evaluating, analysing, comparing or predicting their professional practices, with violations punishable under Articles 226-18, 226-24, and 226-31 of the penal code, carrying up to five years' imprisonment. This appears to be the first such ban anywhere in the world, and it followed directly from an earlier effort to make all French case law freely accessible — which had been seen as a transparency advance. The Constitutional Council upheld it, reasoning that judge-by-judge analytics could facilitate strategies for choosing courts and judges, likely altering the functioning of justice, and that the prohibition would not infringe the right to a fair procedure.

Two lessons. First, states will legislate against judicial analytics when it threatens institutional integrity — capability does not translate automatically into permission. Second, the French move reveals a genuine tension: the same transparency that enables accountability enables forum-shopping and modeling.

5.3 China: maximal deployment, explicit limit

China has gone furthest in judicial AI and still stops short. Since the Supreme People's Court introduced "Smart Courts" in 2017, China has built a system centred on smart adjudication, enforcement, services, and management, and in 2022 launched a nationwide unified Online Service of People's Courts supporting the entire litigation process from filing to judgment and enforcement. Shanghai's "206" system supports criminal decision-making across the procuratorial, trial, and trial-supervision stages, and is directly supported by the Central Political and Legal Affairs Commission — arguably the world's most advanced AI-assisted adjudication deployment.

The SPC's 2022 Opinions on Regulating and Strengthening the Applications of Artificial Intelligence in the Judicial Fields set targets requiring a competent AI system by 2025 and improved rules by 2030 — while underscoring that rulings must always be made by judges, meaning AI cannot replace courts in deciding cases, and that results from the technology serve only as supplemental references in judicial supervision and management.

That the most enthusiastic state adopter, operating without the constraints of an independent judiciary or an entrenched due-process tradition, still refuses to let the machine decide, is strong evidence about the practical limits.

5.4 Estonia: a case study in evidence hygiene

The most widely cited example of a "robot judge" does not exist. A 2019 *Wired* article reported that Estonia's Ministry of Justice had tasked chief data officer Ott Velsberg with designing a robot judge for small claims under €7,000. The story propagated through the World Economic Forum, academic articles, and countless policy papers. In February 2022 the Estonian Ministry of Justice published a flat denial: the article is misleading; there has not been that kind of project or even an ambition in the Estonian public sector; the Ministry does not develop an AI robot judge for small claims or general court procedures to replace the human judge. It is searching for ICT means to make the courts' workload more bearable, and looking at optimization and automation of procedural steps. One peer-reviewed review of European e-justice states the matter bluntly: the example continues to be hailed as one of the most advanced AI projects in European justice systems, and it is simply fake news.

Anyone assessing this field should treat the Estonia episode as a calibration exercise. A substantial share of the "AI is already judging cases" literature traces back to reporting that the responsible ministry denies.

5.5 United States

Constitutional obstacles. Article III vests the judicial power in judges appointed under specified procedures with life tenure. The Sixth and Seventh Amendments guarantee jury trial in criminal cases and in civil suits at common law. Due process requires notice, an opportunity to be heard before a neutral decision-maker, and — in most contexts — a reasoned decision susceptible to appellate review. None of these can be satisfied by a system whose reasoning is a post-hoc rationalization of a computation, and amending them requires constitutional amendment, not statute.

Current institutional posture. The federal judiciary distributed interim AI guidance to courts on July 31, 2025. It aims to allow use of and experimentation with AI tools while preserving the integrity and independence of the federal courts and the judicial process, and addresses procurement, security, confidentiality of judiciary data, and AI education. Reported specifics include directions to review and independently verify AI-generated work product, cautions against delegating core judicial functions to AI, encouragement to define locally approved tasks, and attention to disclosure and confidentiality; a September 2025 Strategic Plan separately called for an AI governance framework.

Actual usage. The first random-sample survey of federal judges, by Anika Jaitley, Daniel Linna, Judge Xavier Rodriguez, V.S. Subrahmanian, and Siyu Tao, published in *The Sedona Conference Journal* Vol. 27, surveyed 502 randomly selected bankruptcy, magistrate, district, and appellate judges in late 2025, with 112 responding (22.3%

response rate). 61.6% of responding judges use at least one AI tool in their judicial work, but only 5.4% report daily use and 17% weekly; 19.6% monthly, 19.6% rarely, and 38.4% never. AI is present in federal chambers but not yet a routine, embedded part of most judges' decision-making. On governance: approximately 20% of judges formally prohibit AI use, 17.6% discourage without formally prohibiting, and 24.1% have no official policy. The response rate warrants caution about non-response bias in either direction.

Evidence law is adapting. The Judicial Conference's Committee on Rules of Practice and Procedure has proposed Federal Rule of Evidence 707, which would subject AI-generated evidence offered at trial without a corresponding expert to expert-testimony-style reliability standards; the public comment period closed February 16, 2026, with the Advisory Committee on Evidence Rules meeting May 7, 2026, and further Rules Enabling Act review to follow.

Fragmentation. State approaches diverge sharply. Illinois permits AI use in court proceedings with ethical safeguards; the Texas Supreme Court has considered a complete ban; Delaware adopted an interim policy that neither prohibits nor requires generative AI. As one federal judge put it in mid-2026, there are no standards right now, and the federal judiciary is operating on an ad hoc basis.

Part VI — A Feasibility Matrix

Synthesizing the above, here is my assessment of what is achievable, on what horizon, and with what confidence.

Function	Technical feasibility	Legal permissibility	Likely by 2035	Desirable?
Legal research, citation retrieval, summarization	High now	Permitted	Near-universal	Yes, with verification duties
Document review, e-discovery, due diligence	High now	Permitted	Already standard	Yes
Drafting under supervision	High now	Permitted	Near-universal	Yes
Case triage, scheduling, docket management, anonymization	High now	Expressly carved out of EU high-risk rules	Widespread	Yes
Self-help guidance for unrepresented litigants	Moderate	Contested (UPL rules)	Widespread where sandboxes permit	Yes, if reliability gated
Automated small-claims / traffic / benefits determination with opt-out to a human	Moderate	Achievable by statute	Plausible in some jurisdictions	Defensible
Sentencing and bail recommendation	Technically easy, normatively fraught	Permitted post-*Loomis*, poorly constrained	Already happening	Doubtful without radical transparency
Contested civil adjudication with disputed facts	Low	Barred in EU; constitutionally fraught in US	Unlikely	No
Credibility determination from live testimony	Very low	Barred	Very unlikely	No
Criminal trial, jury functions, constitutional adjudication	Not a capability question	Requires constitutional amendment in US	No	No

The load-bearing distinction running through this table is between *bounded* decisions — where the applicable rule is determinate, the facts are documentary, the stakes are low, and errors are cheaply reversible — and *unbounded* decisions, which involve contested facts, discretionary standards, novel legal questions, or irreversible consequences. Automation is defensible in the first category and not the second, and the boundary should be drawn by statute rather than by procurement.

Part VII — The Strongest Case for the Other Side

Intellectual honesty requires stating the best version of the opposing argument, which is stronger than most critics allow.

Volokh's test. Eugene Volokh's *Chief Justice Robots* (68 Duke L.J. 1135) makes the sharpest philosophical case. He proposes that if software can create persuasive opinions capable of regularly winning opinion-writing competitions against human judges, and if it can be adequately protected against hacking and similar attacks, we should in principle accept it as a judge — no less reliable and more cost-effective than human judges — even if the opinions do not stem from human judgment. He anticipates the obvious objection about why persuasion rather than correctness: the problem is that there are multiple judgments about what is fair, wise, or correct, and these vary not just among observers — the same observer can often say several positions are defensible. He also anticipates the incrementalist path: some decisions will be more politically palatable for AI judging, at least at first, and AI judging can be tested there before attempts to spread it more broadly.

This argument is difficult to defeat on its own terms. If our only access to judicial quality is through the persuasiveness of written reasons, and a machine produces more persuasive reasons more cheaply, the demand for something further — "real" understanding — needs an account of what that something is and why it matters.

The baseline argument. The status quo is not a well-functioning system being threatened by automation. It is a system in which 92% of the substantial civil legal problems of low-income Americans get no or inadequate help and in which the identity of the randomly assigned adjudicator is often the most consequential fact in an asylum case. Holding automation to a standard of perfection while grading the incumbent on a curve is not analysis; it is status-quo bias. And the consistency evidence is real: LLMs produced significantly lower sentence disparity than 123 retired judges on matched cases.

My response. The persuasion test conflates the epistemic and the political functions of adjudication. A judgment is not merely a claim that a proposition is legally correct; it is an exercise of coercive public authority over a person, and its legitimacy derives partly from the fact that an accountable human being, subject to removal, appeal, professional discipline, and public criticism, took responsibility for it. This is Frank Pasquale's central point in *A Rule of Persons, Not Machines*: to preserve accountability and a humane legal order, reasons must be expressed in language by a responsible person, and this basic requirement for legitimacy limits legal automation. He argues that legal language is essential to realizing rule-of-law principles, and software cannot replicate the ambiguity and interpretive flexibility inherent in human language, warning that legal automation can elide or exclude important human values, necessary improvisations, and irreducibly deliberative governance.

On the consistency point: consistency is a *component* of justice, not a synonym for it. A system that sentences everyone identically wrongly is perfectly consistent. And the accuracy metric in the consistency studies is calibrated to the human average, meaning the machine's demonstrated virtue is variance reduction around an unexamined center.

Part VIII — What Would Have to Be True

Rather than concluding with a verdict, it is more useful to specify falsifiable conditions. Anyone proposing AI adjudication should be required to demonstrate the following, and each is currently unmet:

1. **A resolved fairness specification.** A public, statutory choice of which fairness criterion governs, with explicit acknowledgment of what is sacrificed — because the impossibility theorems guarantee something is.
2. **Auditable models.** No trade-secret adjudicative systems. *Loomis* went the wrong way; the fix is legislative. Independent researchers must have API access at scale, which the Stanford RegLab team identified as currently absent.
3. **Demonstrated groundedness at rates far below current figures.** 17–33% hallucination is not a rounding error in a domain where a fabricated citation is a fabricated legal basis for coercion.
4. **Empirical evidence against automation bias.** Measured departure rates from AI recommendations, with the null hypothesis being that human review is decorative.
5. **Political-valence auditing.** Regular, published, adversarial measurement of the ideological positioning of deployed models, with the fine-tuning corpus and reward model subject to something resembling the scrutiny of a judicial confirmation.
6. **Symmetric access.** Equal-quality tooling for both sides, publicly funded, or the amplification of the repeat-player advantage will swamp any efficiency gain.
7. **Genuine opt-out.** An unconditional right to human adjudication, exercisable without cost or delay penalty — which is what distinguishes automation as an offering from automation as a mandate.
8. **A theory of legal change.** An account of how an automated system could ever overrule its own training distribution.

Conclusion

Can AI replace lawyers? Substantially, yes — and this is already underway, with randomized-trial evidence of significant quality and productivity gains, alongside a rapidly growing docket of sanctions for practitioners who mistook fluency for reliability. The realistic 2035 profile is a legal profession with far fewer hours billed for research and drafting, a far larger volume of assisted self-representation, and a persistent premium on judgment, advocacy, negotiation, and responsibility-bearing.

Can AI replace judges? Not on any timeline visible from here, and the obstacles are not primarily technical. Every serious regulatory regime examined — the EU AI Act, the ECHR framework, the Chinese Supreme People's Court, the U.S. federal judiciary's interim guidance — has independently converged on the same rule: the machine may assist, the human must decide. That convergence across radically different political systems is evidence about the nature of adjudication, not about the state of the technology.

Would AI eliminate political bias and money from justice? This is the part of the proposal that fails most clearly, and it fails on its own terms rather than on mine. Bias cannot be eliminated because the impossibility theorems establish that "unbiased" names a family of mutually incompatible criteria; any system must choose, and the choice is political. Political influence would not be removed but concentrated: the demonstrated left-of-centre tilt of current models and

the demonstrated ease of shifting that tilt by fine-tuning together mean that an automated judiciary's ideology would be a parameter under someone's control. And money would not be removed from justice but redirected — from campaign contributions and billable hours toward procurement contracts, model licensing, and the asymmetric capability advantage of parties who can afford the good version.

The honest reframing is this. AI is a genuinely powerful instrument for reducing the *arbitrariness* of legal outcomes and the *cost* of legal work. It is not an instrument for removing value judgments from law, because there is no such instrument — value judgments in law are not a defect to be engineered out, they are the substance of what is being decided. The choice is never between a political judiciary and an apolitical one. It is between political judgment that is visible, contestable, distributed across thousands of accountable individuals, and correctable through appeal and appointment — and political judgment that is embedded in a specification, deployed at scale, and effectively unreviewable.

That trade may still be worth making for parking tickets. It is not worth making for anything that matters.

References

Empirical studies of AI capability in law

- Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies*, 22(2), 216. https://onlinelibrary.wiley.com/doi/full/10.1111/jels.12413 | Preprint: https://arxiv.org/abs/2405.20362
- Schwarcz, D., Manning, S., Prescott, J. J., Barry, P., Cleveland, D. R., & Rich, B. (2026). AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice. https://journals.sagepub.com/doi/10.1177/2755323X261427048 | SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111
- Martínez, E. (2025). Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, 33(3), 581–604. https://link.springer.com/article/10.1007/s10506-024-09396-9
- Posner, E. A., & Saran, S. (2026). Judge AI: A Case-Study of Large Language Models as Judges. https://journals.sagepub.com/doi/10.1177/2755323X261433614 | Working paper: https://chicagounbound.uchicago.edu/law_and_economics/1044
- Gazal-Ayal, O., Elyoseph, Z., & Solomon, A. (2026). Evaluating Large Language Models as Judicial Decision-Makers. *Justice Quarterly*, 1184–1219. https://www.tandfonline.com/doi/full/10.1080/07418825.2026.2618254
- Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights. *PeerJ Computer Science*, 2:e93. https://doi.org/10.7717/peerj-cs.93
- Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, 28(2), 237–266. https://link.springer.com/article/10.1007/s10506-019-09255-y
- Medvedeva, M., Wieling, M., & Vols, M. (2022). Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*. https://link.springer.com/article/10.1007/s10506-021-09306-3
- Jaitley, A., Linna, D. W., Rodriguez, X., Subrahmanian, V. S., & Tao, S. (2026). Artificial Intelligence in Federal Courts: A Random-Sample Survey of Judges. *The Sedona Conference Journal*, 27. https://www.thesedonaconference.org/publication/Artificial_Intelligence_in_Federal_Courts

- Charlotin, D. AI Hallucination Cases Database. HEC Paris Smart Law Hub. https://www.damiencharlotin.com/hallucinations/ (Running tally of court decisions in which a tribunal found reliance on AI-fabricated material; downloadable in CSV.)

Algorithmic fairness

- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *ITCS 2017*. https://drops.dagstuhl.de/storage/00lipics/lipics-vol067-itcs2017/LIPIcs.ITCS.2017.43/LIPIcs.ITCS.2017.43.pdf
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2).
- Beigang, F. (2023). Reconciling Algorithmic Fairness Criteria. *Philosophy & Public Affairs*. https://onlinelibrary.wiley.com/doi/10.1111/papa.12233

Political valence of language models

- Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE*, 19(7), e0306621. https://doi.org/10.1371/journal.pone.0306621
- Rozado, D. (2025). Measuring Political Preferences in AI Systems: An Integrative Approach. https://arxiv.org/abs/2503.10649

Legal theory and jurisprudence

- Volokh, E. (2019). Chief Justice Robots. *Duke Law Journal*, 68, 1135–1192. https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=3973&context=dlj
- Pasquale, F. (2019). A Rule of Persons, Not Machines: The Limits of Legal Automation. *George Washington Law Review*, 87, 1. https://www.gwlr.org/wp-content/uploads/2019/01/87-Geo.-Wash.-L.-Rev.-1.pdf
- Susskind, R. (2019). *Online Courts and the Future of Justice*. Oxford University Press.
- Chatziathanasiou, K. (2022). Beware the Lure of Narratives: "Hungry Judges" Should Not Motivate the Use of "Artificial Intelligence" in Law. *German Law Journal*, 23(4), 452–464. https://doi.org/10.1017/glj.2022.32

Judicial behaviour, money, and disparity

- Kang, M. S., & Shepherd, J. M. (2011). The Partisan Price of Justice: An Empirical Analysis of Campaign Contributions and Judicial Decisions. *NYU Law Review*, 86. https://www.nyulawreview.org/wp-content/uploads/2018/08/NYULawReview-86-1-Kang-Shepherd.pdf
- Kang, M. S., & Shepherd, J. M. (2023). *Free to Judge: The Power of Campaign Money in Judicial Elections*. Stanford University Press. https://www.sup.org/books/law/free-judge
- Shepherd, J. (2013). Justice at Risk: An Empirical Analysis of Campaign Contributions and Judicial Decisions. https://www.acslaw.org/analysis/reports/justice-at-risk/
- Bonica, A., & Rosenthal, H. (2017). Replication study on judicial campaign finance (Cornell Law faculty publications). https://scholarship.law.cornell.edu/facpub/1654
- Ramji-Nogales, J., Schoenholtz, A. I., & Schrag, P. G. (2007). Refugee Roulette: Disparities in Asylum Adjudication. *Stanford Law Review*, 60. https://www.stanfordlawreview.org/wp-content/uploads/sites/3/2010/04/RefugeeRoulette.pdf
- Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *PNAS*, 108(17), 6889–6892. https://doi.org/10.1073/pnas.1018033108

- Weinshall-Margel, K., & Shapard, J. (2011). Overlooked factors in the analysis of parole decisions. *PNAS*, 108(42), E833. https://pmc.ncbi.nlm.nih.gov/articles/PMC3198355

Access to justice

- Legal Services Corporation (2022). *The Justice Gap: The Unmet Civil Legal Needs of Low-income Americans*. https://justicegap.lsc.gov/
- Legal Services Corporation (2017). *The Justice Gap: Measuring the Unmet Civil Legal Needs of Low-income Americans*. https://www.lsc.gov/initiatives/justice-gap-research
- IAALS (2025). Alternative Business Structures in the U.S.: What We Know and What We Still Need to Learn. https://iaals.du.edu/news/alternative-business-structures-us-what-we-know-and-what-we-still-need-learn
- IAALS. An Interim Evaluation of Utah's Legal Regulatory Sandbox. https://iaals.du.edu/publications/interim-evaluation-utah-process-evaluation
- Stanford Law School (2025). Regulatory Innovation at the Crossroads: Five Years of Data on Entity-Regulation Reform in Arizona and Utah. https://law.stanford.edu/2025/06/02/regulatory-innovation-at-the-crossroads-five-years-of-data-on-entity-regulation-reform-in-arizona-and-utah/
- Chaire LexUM (survey of BC Civil Resolution Tribunal users). https://chairelexum.openum.ca/s/3203

Law, regulation, and cases

- *State v. Loomis*, 2016 WI 68. Case note: 130 *Harvard Law Review* 1530. https://harvardlawreview.org/print/vol-130/state-v-loomis/
- Regulation (EU) 2024/1689 (AI Act), Annex III, point 8(a). https://artificialintelligenceact.eu/annex/3/ | Recital 61: https://ai-act-service-desk.ec.europa.eu/en/ai-act/recital-61
- Judges-in-the-loop? Judicial involvement in human oversight of high-risk decision support systems under the EU AI Act. *International Journal of Law and Information Technology*. https://academic.oup.com/ijlit/article/doi/10.1093/ijlit/eaag001/8475419
- France, Article 33, Loi de programmation 2018-2022 et de réforme pour la justice. Discussion: https://www.jurist.org/news/2019/06/new-france-law-bans-use-of-analytics-to-determine-judge-behavior/
- Supreme People's Court of China (2022). Opinions on Regulating and Strengthening the Applications of Artificial Intelligence in the Judicial Fields. https://english.court.gov.cn/2022-12/12/c_838810.htm
- Estonian Ministry of Justice and Digital Affairs. "Estonia does not develop AI Judge." https://www.justdigi.ee/en/news/estonia-does-not-develop-ai-judge
- Velicogna, M. From Court Automation to e-Justice and Beyond in Europe. *International Journal for Court Administration*. https://iacajournal.org/articles/10.36745/ijca.640
- Administrative Office of the U.S. Courts, interim AI guidance (July 31, 2025). Reported at https://fedscoop.com/interim-ai-guidance-us-courts-aims-experimentation-guardrails/
- Proposed Federal Rule of Evidence 707 (AI-generated evidence). https://www.americanbar.org/advocacy/governmental_legislative_work/publications/washingtonletter/april-26-wl/ai-in-the-courts-update-0426wl/

Short answer: I'd almost certainly pass, probably comfortably. But the honest version of that answer is more interesting than the number, and it undercuts itself in ways that connect directly to the paper you just had me write.

What's actually been measured

No one has published an independent, properly graded Uniform Bar Exam evaluation of a current Claude model. The most recent bar-adjacent data points:

Multiple choice (MBE). A March 2026 controlled benchmark ran ChatGPT 5.2, Claude Opus 4.5, and Gemini 3 Pro against the full NCBE MBE Complete Practice Exam (questions 1–200, no exclusions) under standardized no-web conditions. The three general-purpose systems each missed between 13 and 23 questions. That's roughly 88–93% raw, which converts to an MBE scaled score well above any jurisdiction's threshold. LawSites

The caveats are serious, and the reporting flagged them: the benchmark covers only multiple-choice reasoning, the evaluation was run entirely by the team that built the competing product, scoring used an AI judge from the same model family as one of the systems tested, and each question was run once, so there are no confidence intervals. The authors also concede training data exposure cannot be ruled out for any evaluated system, which for a commercially published practice exam is close to a certainty. LawSitesLawSites

Essays. This is where the picture gets murkier, and where GPT-4's collapse showed up in Martínez's reanalysis (roughly 15th percentile against licensed attorneys). No comparable independent UBE essay grading has been published for current reasoning models.

The best available proxies are foreign bar exams, and they diverge sharply:

Exam	Result	Interpretation
Brazilian OAB (*oab-bench*)	Claude 3.5 Sonnet best at 7.93/10, passing all 21 exams	Comfortable pass
Greek Bar (*GreekBarBench*)	Claude 3.7 Sonnet at 7.71 vs. average legal expert 7.78, passing score 6.00, top-5% of experts 8.87	Passes; matches average expert, not top experts
Japanese Bar writing task	Average 25.6/100 across all answers; only 9.0% rated *Adequate*, 5.3% *Good*, none *Excellent*. Claude 3 Opus best at 37.1% when given curated statutes	Fails badly

The Japanese result is the one worth sitting with. It's a statute-heavy, open-ended, civil-law writing task, and providing the governing statutes helped substantially. That pattern suggests the models are strong at retrieval-plus-pattern-matching and much weaker at sustained doctrinal construction from first principles.

My estimate

For the U.S. UBE specifically: near-ceiling on the MBE, competent but unexceptional on the MEE and MPT, comfortable overall pass. If forced to guess a percentile against first-time takers, somewhere in the 70s–80s, with genuinely wide error bars. I'd bet against a top-decile placement.

I have no evaluation data on the model you're currently talking to. Everything above is Opus 4.5 through 4.7 and earlier.

Three reasons the number doesn't mean what it appears to

Contamination is structural here, not incidental. The bar exam is closed-book, tests majority rules rather than jurisdiction-specific law, and its released materials are among the most heavily reproduced legal texts in existence. This isn't a test of legal reasoning that happens to be in the training corpus. It's a test *of the corpus*.

The bar tests the automatable subset. Issue-spotting from a clean fact pattern with the facts stipulated, no client to interview, no witness whose credibility is contested, no adversary generating new facts, no discretionary remedy to

fashion. It measures exactly what Part II of the paper classified as low-difficulty and omits everything classified as high.

The failure mode is the whole problem. In that MBE benchmark, among 52 incorrect outputs across the general-purpose systems, 49 delivered assertive, well-structured reasoning that gave no signal of uncertainty. A wrong answer that reads like competent lawyering is the expensive kind. This is the same mechanism producing the 1,600-plus sanctioned filings in Charlotin's database, and it is precisely what an exam score cannot detect, because on a graded test a confident wrong answer and a hedged wrong answer score identically. LawSites

One moving-target note: most jurisdictions transition from the MBE/MEE/MPT to the NextGen UBE between July 2026 and July 2028, with Conflicts of Laws, Family Law, Trusts & Estates, Wills, and Secured Transactions dropping off the MEE as of July 2026. So even the benchmark is being redesigned underneath the question. NCBEJD Advising

The framing I'd defend: passing the bar is evidence of eligibility to be supervised, not evidence of judgment. Every jurisdiction already knows this, which is why passing it earns you a license and not a bench.

Let me split this, because "how would you fare" now has two very different answers.

The bar exam becomes uninformative

On the exam itself, 12–24 months out, I'd expect the question to stop being interesting. MBE performance is already 88–93% and the remaining errors are mostly in the tail where question ambiguity and model error are hard to separate. Essays are the growth area, and essay generation is exactly what reasoning models have improved at fastest.

There's also a measurement problem coming. Jurisdictions transition to the NextGen UBE between July 2026 and July 2028, so by late 2027 there won't be a stable exam to compare against, and any new released materials will be freshly uncontaminated for maybe one cycle before they enter the corpus. My guess is that by 2028 nobody serious cites bar scores as evidence of anything.

The trend lines that actually matter

METR's time-horizon work is the most useful quantitative anchor. The finding: the length of task (measured by how long a human expert takes) that a model completes autonomously at 50% reliability has been doubling roughly every seven months for six years. Their updated methodology puts the post-2023 doubling time at 131 days, about 20% faster. Over 2024–2025 specifically, the doubling was closer to every four months, and there's no visible plateau. Claude Opus 4.6 measured around a 12-hour 50% horizon.

Naively extrapolated over 24 months, that's 3–6 doublings: somewhere between four and thirty days of expert-equivalent work.

Here's why that number is misleading, and it's the single most important caveat in this whole answer. METR also measures the 80% horizon, and it's 4–6x shorter than the 50% horizon, growing at essentially the same rate. So the useful translation is: **what a model can do unreliably today, it does reliably in about 14 months.** Reliability lags capability by roughly two doublings, permanently.

The real-work benchmark tells a compatible story. On GDPval, which uses actual professional deliverables graded by blinded expert pairwise comparison across 44 occupations including lawyers, Claude Opus 4.1 was rated as good as or better than industry experts on 49% of tasks, with GPT-5-high at 40.6%. OpenAI noted the Claude figure may partly reflect a preference for its graphics rather than substance. Tasks averaged about 7 hours of expert time. Performance more than doubled in a year.

The infrastructure question

You specifically asked about infrastructure, and I think the answer there is counterintuitive: the constraints are real but they mostly bite *after* your window.

Compute stock at frontier labs has grown around 2.2x per year, and training compute 4–5x per year. Epoch argues this must slow, because each 10x increase in compute stock adds roughly a year of lead time, and because training-run duration is approaching an economic ceiling. Duration has grown 1.4x per year and runs beyond about nine months are sub-optimal, a limit likely hit around 2027. Power is now the binding constraint rather than chips: Gartner projects 40% of AI data centers power-constrained by 2027, with grid approval timelines of 24–36 months in major markets. Frontier training campuses draw 100–150 MW today against projections of 1–2 GW by 2028.

But Epoch's own qualifier matters here: even with lengthening lead times, labs can sustain roughly 5x annual training compute growth for another one to two years by reallocating a larger share of existing compute to training. The capital and buildout covering the next 24 months is largely already committed. So the 2027–2028 window is probably the last stretch running on momentum already purchased, with any deceleration showing up in 2029+ model generations.

My actual forecast for legal work

Capability	Sept 2027 – Sept 2028	Confidence
Research, summarization, discovery, first-draft generation	Fully routine, verification tooling standard	High
Multi-hour autonomous workflows (full privilege review, complete brief with verified citations, deposition prep)	Reliable at supervised-associate level	Moderate-high
Multi-day autonomous matter handling	Achievable at 50%, not at 80%	Moderate
Fabrication rate in RAG legal products	Materially below the 17–33% Magesh et al. found, but not zero	Moderate
Anything in the paper's "unbounded" category	Essentially unchanged	High

The thing I'd bet on hardest is that **the reliability gap closes faster than the capability frontier advances, and that this is what actually changes the profession.** Today's failure mode isn't that models can't draft a brief. It's that among incorrect outputs in that MBE benchmark, 49 of 52 were confidently wrong with no uncertainty signal. If calibration improves — models that reliably flag "I am not sure this case says what I think it says" — that does more for real legal deployment than another 10% on any benchmark.

What I'd bet against: the regulatory floor moving. The EU AI Act's Recital 61 requirement that final decision-making remain human-driven, the ECHR Article 6 constraints, the U.S. constitutional structure — none of these are on a capability schedule. A 2028 model that comfortably out-drafts a magistrate still won't be permitted to be one.

Three things I'd watch as falsifiers

1. **Charlotin's curve.** It's currently steepening, roughly eight new sanctioned filings a day. If it bends down while AI adoption keeps rising, verification has been solved in practice. If it keeps climbing, capability gains aren't translating into reliability.
2. **An independent replication of the Stanford RegLab study.** If someone reproduces that methodology and finds single-digit hallucination rates, that's the strongest possible signal.
3. **GDPval legal win rates crossing 50% under blinded expert review** with the aesthetic confound controlled.

One honest note on epistemics. You're asking me to forecast my own successors, which is a strange position to be in. I have no privileged access to Anthropic's roadmap, models are known to be poorly calibrated about their own

capabilities, and there's an obvious direction for bias here. I've tried to lean on external measurement rather than introspection for exactly that reason, but treat the confidence column above as softer than it looks.