Online ISSN: 3104-9338
Print ISSN: 3104-932X

From Benchmarks to Courtrooms: A Capability–Reliability–Accountability Survey of Large Language Models in Legal Practice

Yuying Zhang,Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: August 23 2026

Revised: August 30, 2026

Accepted: August 30, 2026

Published online: August 30, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 9 (September 2026)

* Corresponding Author: Xinyu Song (songxinyu@guc.edu.cn)

Abstract. Large language models have entered legal practice rapidly, yet systems that score highly on legal benchmarks have fabricated judicial citations in filed briefs—a contradiction that resource-inventory surveys do not explain. This survey reviews 25 core sources (2019–2026, plus the 1987–2003 symbolic genealogy) through an original capability–reliability–accountability (CRA) framework; reported figures are quoted only from peer-reviewed studies, cited case law, or first-party technical reports. Measured capability has matured across three benchmark generations, with frontier models exceeding 85% on aggregate legal-reasoning tasks. Reliability has not kept pace: benchmark validity limits, citation hallucination rates of 17–43% even under retrieval grounding, and pipeline-level retrieval failures independently sever the link between scores and trustworthiness. Accountability responses—high-risk regulation, neuro-symbolic scaffolding descending from the HYPO tradition, and rubric-bound LLM-as-a-judge evaluation—converge on a single strategy: re-embedding statistical systems in inspectable structure. The CRA framework explains the benchmark-to-courtroom gap, positions the field's open problems, and transfers to other high-risk professional domains.

Keywords:*hallucination; retrieval-augmented generation; judgment prediction; neuro-symbolic reasoning; evaluation validity*

1. Introduction

1.1. Background and Motivation

Large language models (LLMs)—neural networks trained on web-scale text to predict and generate language—have moved into legal practice faster than almost any other professional

domain. General-purpose systems now draft contracts, summarize depositions, compare precedent, and answer research queries in natural language, while specialized offerings from major legal publishers—Thomson Reuters' CoCounsel, LexisNexis' Lexis+ AI, and the independent platform Harvey—connect the same underlying models to proprietary corpora of cases, statutes, and firm documents [1]. Courts and bar associations have responded with guidance on disclosure and verification rather than prohibition, treating adoption as a matter of time. For a profession whose raw material is authoritative text, the appeal is obvious; so is the exposure.

The fit between the technology and the profession explains the speed, and understanding the fit is prerequisite to understanding the failure. Legal work is disproportionately text transformation under constraint: extracting issues from records, locating authorities that match fact patterns, assembling citations and quotations into structured documents, checking that propositions say what they are cited for. These are the operations language models perform most fluently, and the domain supplies what general-purpose applications lack—massive, stable, authoritatively organized text corpora (reporters, codes, registries) against which outputs can be checked and from which systems can be grounded [1]. The same properties explain the form the failures take. A profession organized around citation inherits hallucination as its signature risk: the single most consequential thing a legal AI can do is assert that an authority exists and says what the drafter needs it to say, and that assertion is exactly where generative fluency and professional validity come apart. The domain's own epistemology—authority must be located, verified, and characterized accurately—turns out to be a checklist purpose-built for detecting model failure, which is why the legal domain produces the field's clearest hallucination numbers rather than despite being text-native, but because of it [1]. The technology's fit with the work and the work's fit with auditing the technology are, in other words, the same properties read from two directions.

The stakes of this transition are captured by two events from the same year. In 2023, two New York attorneys were sanctioned five thousand dollars in Mata v. Avianca for filing a brief that cited multiple non-existent judicial opinions produced by ChatGPT—complete with invented quotations and internal citations [2]. In the same year, the LegalBench collaboration reported that twenty open-source and commercial LLMs exhibited substantial, measurable legal reasoning ability across 162 curated tasks [3]. The contradiction is pointed: systems that score well on legal benchmarks can still fabricate the very authorities on which legal argument depends.

This survey argues that the contradiction is best understood not as a property of any single model but as a structural gap between three questions the field must answer separately: what legal LLMs can do (measured capability), whether their outputs can be trusted (reliability), and who answers for them when they fail (accountability). Existing surveys, reviewed in Section 1.3, largely organize the field as an inventory of models, datasets, and frameworks; they do not treat the relationship among these three questions as an object of analysis. We do.

The three axes earn their independence by failing independently, which is what makes the framework diagnostic rather than decorative. A system's measured capability places no ceiling under its reliability: the same model that tops a reasoning benchmark can fabricate the citation its answer rests on, because benchmark and fabrication are different properties of different acts—one of answering well on curated tasks, one of asserting authority that exists. Reliability, similarly, does not entail accountability: a system whose outputs are verifiably correct ninety-two times out of a hundred is precisely the system whose twentieth failure names no responsible party unless some structure—professional, contractual, or regulatory—has been built to carry it. And accountability does not purchase capability or reliability; it purchases allocation of whatever risk remains. The legal profession has lived inside exactly this structure for centuries: competence (what a lawyer can do), diligence (whether the work is right), and liability (who answers when it is not) are separate doctrines precisely because they fail separately, and the bar's existing guidance treats AI outputs as raising all three at once. The CRA framework is, in that sense, not imported into law from machine learning but recovered from law's own professional architecture—an observation that explains both why the framework's recommendations in Section 4 sound familiar to lawyers, and why purely technical treatments of the benchmark-to-courtroom gap have repeatedly mispredicted the profession's responses. Tracking the field along the capability–reliability–accountability (CRA) axis, we show that (i) capability measurement has matured rapidly and now extends from bare models to deployed pipelines; (ii) reliability has not kept pace, because benchmark validity limits, hallucination, and retrieval-pipeline failure are independent failure sources that aggregate scores do not separate; and (iii) accountability mechanisms—risk-based regulation, revived symbolic reasoning structures, and structured meta-evaluation—are converging on a single strategy: re-embedding statistical systems in inspectable structure.

1.2. Survey Methodology

We conducted a targeted literature search across arXiv, the ACL Anthology, NeurIPS and ICAIL proceedings, the journal Artificial Intelligence and Law, and primary legal materials

(statutes and reported cases). The search covered 2019–2026 for technical work, with older symbolic AI and argumentation literature (1987–2003) included selectively where it clarifies the genealogy of current systems. Query families combined legal-task terms (e.g., judgment prediction, legal reasoning, contract analysis), method terms (e.g., large language model, retrieval-augmented generation, test-time scaling), and evaluation terms (e.g., hallucination, benchmark, LLM-as-a-judge). Candidates were retained if they either reported empirical results on legal tasks, introduced a benchmark or evaluation methodology, or carried argumentative weight for the reliability or accountability discussion; purely speculative position pieces were excluded. Reported numbers are quoted only from peer-reviewed publications, cited court opinions, or first-party technical reports—never from secondary commentary. This procedure yielded the 25 core sources analyzed below, supplemented by industry technical reports where production behavior, rather than peer-reviewed claims, is the object of study.

Two evidentiary rules governed the survey's assembly, and stating them lets readers weigh its claims. First, primary-source discipline: every performance figure, hallucination rate, and benchmark result is quoted from the study that produced it, never from another survey's summary of it; every litigation fact is taken from the court's own order or the complaint; every product capability is attributed to the vendor's first-party documentation, with the vendor's interest explicitly noted when relevant. Second, verifiability discipline: each source's author list, title, and venue were confirmed against its primary record before inclusion, and sources that could not be verified with confidence were dropped rather than cited on recollection—a policy that occasionally cost the survey a convenient citation but that it treats as non-negotiable for work whose subject is precisely the difference between checked and unchecked assertion. A survey of legal AI that fabricates or misremembers its sources would refute itself; the two rules are how the survey avoids doing so, and they are the same rules, transposed into evaluation practice, that Sections 3 and 4 argue the deployed systems must adopt.

1.3. Positioning Against Existing Surveys

Two recent surveys frame the field's current self-understanding. Hou et al. catalog 16 legal LLM families, 47 LLM-based frameworks, 15 benchmarks, and 29 datasets, providing the most complete resource inventory to date [4]. Dehghani offers a broader overview of LLMs within legal systems, written for a social-science readership and attentive to institutional context [5]. Both are organized around artifacts—what exists—rather than around the evidential status of the claims made about those artifacts. Neither, for example, connects benchmark scores to the documented failure rates of the deployed systems those benchmarks are supposed to certify.

Table 1 contrasts their coverage with this survey's.

Table 1. Coverage comparison between recent legal-AI surveys and this work.

Dimension	Hou et al. [4]	Dehghani [5]	This survey
Measured capability	Resource inventory (16 LLM families, 15 benchmarks)	General overview	Benchmark lineage 2019–2025 and the capability boundary
Reliability	Not a focus	Brief	Validity critique, citation hallucination, deployment gap
Accountability	Brief	Brief	Regulation, neuro-symbolic return, judge meta-evaluation

The remainder of the paper follows the CRA structure. Section 2 reviews what benchmarks can legitimately measure and where the capability boundary lies. Section 3 examines three independent reasons why measured capability does not translate into deployed reliability. Section 4 analyzes the accountability responses that are closing the gap. Section 5 concludes.

2. Measured Capability: What Legal LLMs Can Do

2.1. From LEGAL-BERT to Legal LLMs

The modern capability story begins with domain-adaptive pretraining. LEGAL-BERT demonstrated that continued pretraining on English legal corpora—EU legislation, English case law, and contracts—improved performance on legal NLP tasks relative to general BERT [6], and the recipe was quickly replicated in other jurisdictions, most prominently for Italian legal text [7]. In parallel, the first systematic neural judgment-prediction study on European Court of Human Rights cases established legal judgment prediction (LJP) as the signature benchmark task of the transformer era: given the facts of a case, predict the court's outcome [8]. These models were small by current standards, but they fixed the field's basic template—take a general-purpose architecture, expose it to legal text, and measure task-specific gains—and they surfaced a problem that would prove durable: when the facts section of a judgment already telegraphs its outcome, a high-accuracy model may be reading cues rather than doing law.

The cue-reading suspicion deserves elaboration because it was the field's first validity crisis and its lesson recurs at every later generation. The European Court of Human Rights publishes judgments in a conventionally ordered form—facts, then the court's legal analysis, then outcome—so a model trained on full judgments can reach high accuracy by exploiting the rhetorical regularities of the genre: language in the facts section that signals how the court will

eventually weigh it. The 2019 judgment-prediction study was scrupulous about this, testing not only full-text models but fact-only variants and ablations designed to separate genuine prediction from artifact exploitation, and its headline numbers carried an explicit warning that models capture "surface information" rather than case reasoning [8]. That warning proved prophetic: every subsequent leap in legal-benchmark scores has been accompanied by the same methodological anxiety at a new scale—the bar-exam analyses that followed had to argue about answer-order artifacts and memorized model answers, and the reasoning-benchmark era inherited the question of whether chain-of-thought improvements reflect legal inference or better pattern completion [9]. The pattern across generations is stable enough to state as a rule: a legal benchmark's score is valid exactly to the degree its construction anticipates the shortcuts a language model will find. LegalBench's design—the next subsection's subject—was the first systematic attempt to build that anticipation into a benchmark's architecture [3].

The template changed qualitatively with instruction-tuned LLMs. Rather than specializing a model, practitioners now specialize a prompt, a retrieval corpus, or both. The difference matters for measurement. In the LEGAL-BERT era, capability gains were attributable to legal pretraining; in the LLM era, measured capability reflects an interaction among model scale, instruction tuning, and task elicitation—chain-of-thought prompting alone, with no legal training at all, shifts reasoning benchmarks substantially [9]. Attribution of capability to "the model" therefore became less straightforward precisely as headline scores improved, and the interpretive weight carried by a benchmark number quietly migrated from the artifact to the measurement arrangement around it. That migration sets up the central question of this survey: if a benchmark score measures an arrangement rather than a model, what does it license practitioners to conclude? Sections 3 and 4 give the answer in two parts—less than it appears (Section 3), and less than regulation will accept (Section 4).

2.2. Benchmarks and What They Measure

Benchmarks operationalize capability claims, and the legal domain has developed them in three generations of increasing fidelity. LexGLUE consolidated seven English legal NLU datasets—question answering, entailment, named-entity recognition, and related tasks—into a single evaluation suite, serving as the field's analogue of general-language benchmarks [10]. LegalBench, built collaboratively by 40 legal academics, marked the second generation: rather than importing NLP task types, it decomposes legal reasoning itself into 162 tasks spanning six reasoning categories aligned with the classic issue–rule–application–conclusion (IRAC) structure used in legal education, and reports evaluations of 20 open-source and commercial

models [3]. Its tasks range from issue spotting and rule recall—where models now perform near ceiling—to applying a negotiated contract term to an unanticipated fact pattern, where performance degrades markedly.

LegalBench's construction method is as much a contribution as its scores, because it attacked the cue-reading problem architecturally rather than statistically. Rather than a single large exam, it aggregated 162 small, targeted tasks contributed by forty collaborating legal academics—each task isolating one reasoning primitive (identifying the issue presented by a fact pattern, extracting the rule a court applied, comparing holdings for relevant similarity) and formatted in the structure the contributor judged most faithful to the underlying skill [3]. The decomposition has two properties generic benchmarks lack. It localizes failure: a model that answers well on rule recall but poorly on application is not "worse overall" but diagnosably strong and weak at particular professional operations, which is information a benchmark-as-exam cannot produce. And it distributes validity across many task authors rather than one benchmark committee, making a single exploit that inflates the aggregate score correspondingly harder—each task arrives with its own construction rationale, its own intended shortcut-resistance [3]. The IRAC alignment is the method's legal-harness: by mapping the 162 tasks onto the issue–rule–application–conclusion skeleton, the benchmark makes scores interpretable against the profession's own account of what reasoning is, so that a score report reads not as "the model knows law" but as a profile of which professional operations it performs—exactly the form of evidence a supervised deployment decision requires [3]. LawBench extends measurement to Chinese law with a task suite organized around three capability tiers: recalling legal knowledge, understanding it, and applying it [11]. The third generation moves from models to systems: Harvey's BigLaw Bench evaluates retrieval-augmented production pipelines on tasks drawn from associates' actual work, treating the retrieval layer as a first-class object of measurement [1]. Table 2 summarizes this lineage.

Table 2. Major evaluation efforts for legal LLMs.

Benchmark	Year	Scope	Language	Primary focus
LexGLUE [10]	2022	7 datasets	English	Legal language understanding
LegalBench [3]	2023	162 tasks, 6 reasoning types	English	Legal reasoning (IRAC-aligned)
LawBench [11]	2023	Three-tier capability suite	Chinese	Legal knowledge and application
BigLaw Bench [1]	2025	Production RAG tasks	English	Deployed pipeline quality

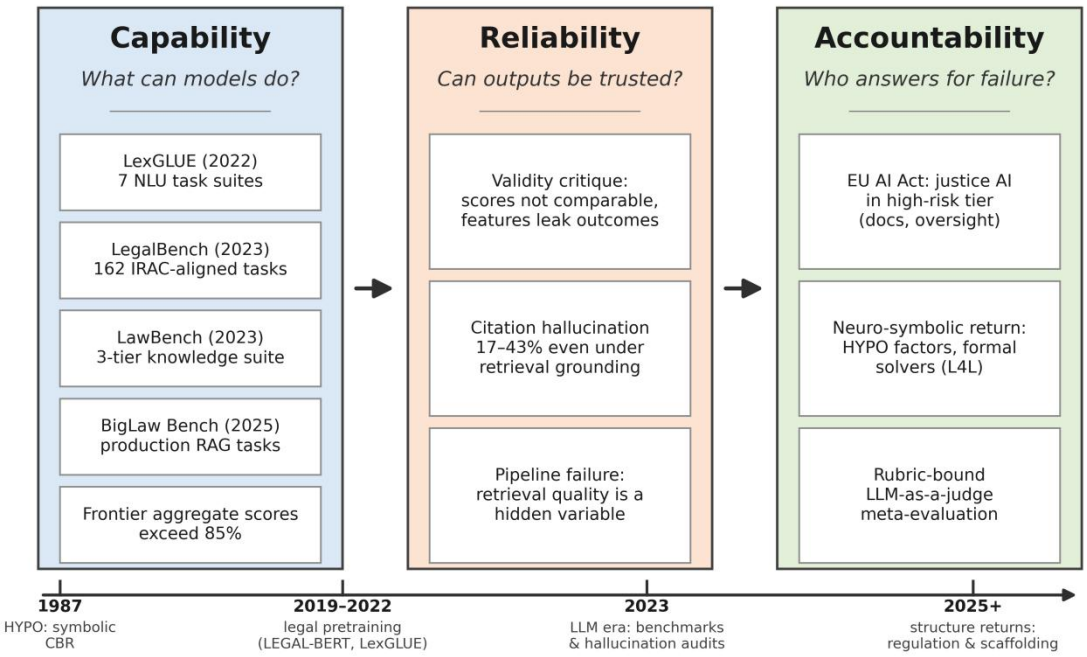

Figure 1. The capability–reliability–accountability (CRA) framework. Left: three generations of capability benchmarks. Middle: independent reliability failure sources. Right: accountability mechanisms that re-impose structure. Bottom: the field's timeline.

Figure 1 situates this progression within the CRA framework. Scores have risen accordingly: current frontier models exceed 85% on LegalBench's aggregate metric [3], a level that, taken alone, would suggest near-expert performance. The next two subsections explain why it does not.

2.3. The Capability Boundary

Three lines of evidence delimit what aggregate scores mean. First, exam-level studies situate model performance relative to trained humans: ChatGPT earned a low but passing grade (C+ range) on law-school examinations across constitutional law, taxation, and torts [12]; earlier analysis projected GPT-3.5 at a passing range on the Uniform Bar Exam [13]; and OpenAI reported top-tier bar-exam performance for GPT-4 [14]. Yet each of these measures closed-book performance on standardized prompts under examination conditions—precisely the setting law schools use precisely because it is not practice. Second, reasoning-model evaluations: the first systematic study of test-time-scaling models—12 systems, including OpenAI's o1 and DeepSeek-R1, across 17 Chinese and English legal tasks—found that extended inference-time reasoning improves some measures but does not remove outdated legal knowledge, limited statutory-interpretation capacity, or susceptibility to factual hallucination

[15]. More inference, in other words, is not more legal reliability.

Read together, the three lines of evidence also explain why the capability boundary falls where it does, and the pattern is more instructive than any single number. The exam studies [12,13,14] measure closed-book recall under standardized elicitation—the task type whose answers are already in the training distribution; reasoning-model evaluations [15] measure inference-time elaboration—which amplifies whatever the model retrieved rather than correcting it; and task-decomposition studies [3] localize the residual weakness precisely at the step—rule application—whose correct execution depends on facts outside the text: what the tribunal will find, how a jurisdiction weighs a factor, what a clause means against a negotiation history. The boundary, in other words, is not a knowledge frontier but a situatedness frontier: models perform well exactly where the answer is recoverable from authoritative text and degrade exactly where the answer depends on the professional's judgment about context the text does not contain. That placement matters because it predicts which deployments are safe—which is assistive work on text (drafting, extraction, issue spotting) rather than autonomous work on judgment—and the prediction has held in practice: production legal AI concentrates, so far, in the former category [1]. The boundary literature, read as a body, is thus less a catalogue of limitations than a map of the deployment envelope the profession is actually living inside. Third, task decomposition: within IRAC-aligned benchmarks, the rule-application step—linking a selected rule to contested facts—remains the weakest component for even the strongest models [3], which is exactly the step that distinguishes competent legal work from retrieval.

The capability literature thus supports a bounded conclusion. Legal LLMs reliably perform component tasks—issue spotting, rule recall, document classification—at levels that justify deployment as assistive tools under professional supervision, while aggregate benchmark scores systematically overstate readiness for unsupervised legal work. Why they overstate it is the subject of the next section.

3. The Reliability Gap: From Benchmark Scores to Deployed Systems

3.1. The Evaluation Validity Critique

The first source of the gap is metrological: a benchmark score is a claim about a dataset, a metric, and an elicitation procedure—not about courtroom performance. Medvedeva, Wieling, and Vols audited the automatic court-decision-prediction literature against explicit quality

criteria covering data availability, methodological rigor, and reproducibility, and concluded that reported accuracy figures cannot be meaningfully compared across studies and frequently overstate real-world capability [16]. Their audit identified recurring defects that have outlived the pre-LLM models they studied: leakage of the final decision into "predictive" features (a facts section written by the same court that decided the case); unrepresentative samples (over-represented published opinions, excluded settlements); and missing metadata (procedural posture, claim type) that a human lawyer would treat as dispositive. The critique transfers directly to the LLM era because the underlying inferential move is unchanged—extrapolating from a curated static dataset to open-world professional performance.

Each defect in the audit deserves its modern translation, because the LLM era has modernized the surface while preserving the structure. Decision-leakage—then, a facts section written by the deciding court—reappears now as ground-truth contamination: benchmarks whose labels derive from the same professional products the model is asked to generate, so that the model's stylistic fluency is scored as predictive accuracy. Unrepresentative samples—then, published opinions over-represented because databases index them—reappear as benchmark skew: legal LLM training and evaluation concentrate on the well-digitized jurisdictions, so a model's measured competence is competence in a data-rich minority of the world's legal systems. Missing metadata—then, procedural posture absent from case files—reappears as context elision: evaluation prompts that withhold precisely the facts a practitioner would consider dispositive (procedural stage, burden of proof, forum), which guarantees that measured performance prices in guesses a lawyer would never make [16]. The audit's summary statistic—reported accuracies that cannot be compared across studies and that systematically overstate real-world capability—was written about pre-LLM court-decision prediction but describes the leaderboard economy accurately; what has changed is only that the numbers now circulate faster and market better [16]. Benchmark designers are aware of the limit: LegalBench frames itself as measuring the ability to perform legal reasoning tasks, not readiness for practice [3]. The problem is downstream: leaderboard numbers circulate detached from that caveat, and a 90% figure quoted without its scope conditions functions as a safety claim the benchmark never made.

3.2. Hallucination and Fabricated Citations

The second source is generation fidelity, and legal text gives it a distinctive shape. In general-domain use, hallucination is an error; in legal use, a fabricated case is a false representation of the legal record—an invention that misstates what the law is, and that the adversarial system is

structured to catch only if opposing counsel happens to look. Mata v. Avianca made this failure mode emblematic: the court sanctioned the attorneys under Rule 11 for submitting six non-existent opinions, finding that they had abandoned their verification obligations when the tool's output matched their expectations [2]. The episode matters for measurement because citation hallucination is also the most checkable failure mode—asserted authorities can be mechanically validated against case-law databases. That checkability is why Avianca functions as the domain's measurement Rosetta stone: it converts an abstract model deficiency into a documented professional catastrophe with a docket number, and it exhibits the full failure sequence—generation, verification skipped, expectation bias, filing, sanctions—that later quantitative studies would decompose [2]. The court's finding on verification deserves particular emphasis as an evaluation datum: the sanctioned attorneys did use the tool and did ask it to confirm its own citations, and the fabrication survived because asking the source to verify itself is not verification—a methodological error so natural that the audit studies treat it as the default user behavior to be designed against [2]. Professional expectation completed the failure: the outputs looked like case law, cited like case law, and carried the confident register that practitioners read as a reliability signal. Every element of that sequence is now measurable—existence checks, source-independence checks, confidence-calibration studies—which is precisely what makes legal hallucination the best-instrumented reliability failure in applied NLP rather than the least controlled [2].

When they are validated, the results are sobering. The landmark audit of legal research tools found hallucination rates of 17–33% on general legal queries even for retrieval-grounded commercial systems (Lexis+ AI, Westlaw's AI-assisted research) and 43% for GPT-4, refuting vendor claims that grounding eliminates fabrication [17]. Fabrication concentrated precisely where practitioners lean hardest on the tools—specific cases, quotations, and statutes—rather than in open-ended exposition.

The audit's design gives its numbers their evidentiary weight, and its details are worth reporting because they have become the template for subsequent hallucination measurement. Dahl et al. built a fixed battery of 207 prompting tasks spanning open-ended synthesis, specific case lookup, quotation, and statutory research; ran it identically against general-purpose models and the two leading vendor legal-research tools; and scored every output against the Westlaw and LexisNexis databases themselves, distinguishing fabricated authority, miscited authority, and incomplete responses [17]. The headline result inverted the vendor safety claims: grounding reduced but did not eliminate fabrication (17–33% for the commercial tools), while the

ungrounded frontier model fabricated at 43%—and, most consequentially, the distribution of failure tracked task specificity, so the everyday workflows of finding and citing authority were the most contaminated [17]. The study also documented the failure modes users cannot self-detect: miscitation of real cases to wrong propositions produces output indistinguishable from correct authority without reading the underlying opinion, and quotation tasks returned real-sounding passages whose wording no database contained. As a measurement contribution, the audit established the field's minimum standard—database-verified scoring, fixed task battery, vendor-product inclusion—that any subsequent reliability claim about legal RAG must now meet [17]. Legal hallucination is thus doubly distinctive: it is professionally catastrophic at rates far above general-domain benchmarks, and its burden of detection falls on a licensed professional who, as Avianca shows, may reasonably expect that a leading product will not invent authority.

3.3. Retrieval-Augmented Generation and Industry Benchmarks

Retrieval-augmented generation (RAG), introduced to ground model generation in documents retrieved from an external corpus [18], is the dominant architectural response to hallucination: every major legal AI product now connects its models to authoritative content through some RAG variant [1]. The audit evidence above shows what grounding does and does not buy. It narrows the fabrication gap by supplying checkable sources; it does not close it, because the model can still mischaracterize a retrieved case—citing real authority for a proposition it does not support—and because grounding introduces its own failure surface. Retrieval quality becomes a hidden variable between the benchmarked model and the deployed system: a pipeline is only as reliable as its index, its corpus boundaries, and its chunking decisions, none of which appear in any model benchmark.

This is precisely the measurement shift that industry benchmarks register. BigLaw Bench evaluates the pipeline—retrieval over proprietary corpora plus generation—rather than the bare model, and reports retrieval as a first-class failure source alongside generation [1]. The pattern generalizes: consumer benchmarks ask what the model knows; production benchmarks ask what this system, on this corpus, under this workflow, actually delivers. The two questions have different answers, and the distance between them is a market signal: firms with the most at stake are building private evaluations because public ones do not measure the thing they are buying. The same asymmetry suggests what a public reliability infrastructure for legal AI would need to look like: shared, versioned legal corpora; published retrieval diagnostics; and hallucination reporting standardized on citation-level checks—metrology, not leaderboards.

BigLaw Bench's retrieval-first design merits unpacking, because it marks the methodological shift this section has been tracing. The benchmark's tasks are drawn from the firm's own matter workflows, its scoring is anchored to associates' completed work products, and—decisively—its instrumentation separates retrieval failures (the corpus did not contain, or the index did not surface, the controlling document) from generation failures (the right document was retrieved and then mischaracterized) rather than merging them into an end-to-end accuracy figure [1]. The separation is what makes the benchmark diagnostic rather than merely evaluative: a pipeline scoring poorly on retrieval needs corpus and indexing work; one scoring poorly on generation needs grounding and prompting work—and the two remedies have different costs, owners, and failure recurrence patterns. It also quietly rebalances professional responsibility: retrieval quality is a configuration the deploying institution controls, while generation quality remains a property of the licensed model, so the benchmark's structure maps the failure surface onto the parties who can actually fix each part [1]. Generalized, the design sketches the reliability-reporting standard this survey's conclusion will call for: any deployed legal AI should publish its pipeline-stage diagnostics, because "the system is 85% accurate" is a marketing number while "retrieval recall is 90%, generation faithfulness on retrieved documents is 94%" is an operations document [1].

Taken together, the three sources of the gap—invalid extrapolation from benchmarks, residual fabrication under grounding, and pipeline-level failure modes—explain why high aggregate scores and sanctioned briefs can coexist. They also explain the responses surveyed next: each is an attempt to make reliability legible where capability benchmarks cannot.

4. Restoring Structure: Accountability and the Return of Symbolic Scaffolding

The responses to the reliability gap share a single logic. Statistical systems are being re-embedded in structures—legal, formal, and evaluative—that convert opaque capability into inspectable behavior. We examine three instances: governance frameworks, neuro-symbolic architectures, and structured meta-evaluation. Their common premise is that the reliability gap is not closed by more capability but by making what capability exists accountable to external criteria.

4.1. Governance as External Structure

The EU Artificial Intelligence Act is the clearest external imposition of structure. Its risk-based classification places AI systems used in the administration of justice in the high-risk tier,

triggering obligations for risk management, data governance, technical documentation, human oversight, and conformity assessment [19]. For legal AI vendors, the effect is to make reliability properties into auditable artifacts—documented systems, logged oversight decisions, assessable conformity—rather than implicit properties of a model. Professional self-regulation points in the same direction: court policies and bar guidance now require attorney review and, in some jurisdictions, disclosure of AI-generated content, converting residual model error into a professionally allocated duty [2]. Governance of this kind does not measure capability, and it cannot; what it does is assign the unmeasured residue to an accountable party, changing the question from "can the model be trusted?" to "who certifies, documents, and answers for it?"

The Act's obligation set deserves itemization because its items are, one for one, the artifacts this survey's reliability section found missing. Risk management in the high-risk tier requires a documented, continuously maintained process for identifying and mitigating foreseeable risks—the standing counterpart of Section 3's failure catalogue; data governance obligations require training, validation, and testing sets that are relevant, representative, and examined for bias—the statutory form of the validity critique; technical documentation and record-keeping make the system's construction and behavior reconstructible after the fact, which is the precondition for every audit; and human-oversight requirements demand that oversight be exercised by competent persons with genuine authority to intervene—legislation codifying, in effect, that the professional remains the unit of accountability [19]. For legal AI the mapping is unusually direct because the regulated application—administration of justice—is named in the Act's high-risk annex, and because each obligation corresponds to a verification practice the profession already recognizes from its own conduct rules: competence, diligence, documentation, supervision. The Act does not solve the reliability gap; it prices it—converting undocumented reliability claims from a competitive virtue into a compliance liability, and thereby giving vendors the economic incentive the pure methodology literature never could [19].

4.2. Neuro-Symbolic Scaffolding: Classical Legal Reasoning Returns

The second response is architectural, and it is a return. The symbolic era of AI and law built explicit structures for exactly the step where LLMs remain weakest. HYPO modeled adversarial case-based reasoning over legally salient fact dimensions—"factors" that strengthen one side or the other in trade-secret law—generating arguments by moving among a database of precedents [20]. Its successors carried the program from argument construction toward prediction: SMILE extended factor-based case representation to outcome prediction,

demonstrating that structured case comparison and statistical learning could be combined a generation before LLMs made such combinations urgent again [21], while argumentation theory formalized defeasible reasoning through schemes and value preferences [22]. The LLM era initially displaced these structures as end-to-end generation made them seem unnecessary; the reliability gap has brought them back as scaffolding.

The lineage's persistence through the statistical era is the part of the story standard narratives omit, and it matters because it shows the symbolic program was never refuted—only outpaced. HYPO's factor model made a specific, falsifiable claim: that in a bounded doctrine (trade secrets), argument moves on a named set of legally salient dimensions, so the moving itself can be computed—and the system generated three-move arguments among precedents rated on those dimensions, with the dimensions, not the text, carrying the normative analysis [20]. SMILE then demonstrated the claim's extension to prediction: rating cases on the same factors and letting a simple model score argument strength predicted outcomes above baseline, twenty years before "LLM-plus-structured-scoring" reappeared as an architectural recommendation [21]. Argumentation theory contributed the formal semantics—defeasible schemes and value-orderings under which an argument's strength is a computable function of stated premises and preferences [22]. Each strand solves a piece of the rule-application problem that Section 2.3 identified as the models' weakness: HYPO the structure of precedential comparison, SMILE the aggregation of factor strength, argumentation frameworks the defeasibility that flat generation flattens. The neuro-symbolic present is thus not a revival but a recombination—one whose novelty is mostly the language model's new role as the interface between unrestricted professional text and the symbolic layer that never stopped existing [20-22].

The clearest example is the L4L framework, which couples LLM agents with formal solvers so that interpretive choices in statutory reasoning are made explicit and machine-checkable rather than absorbed silently into generation; its design premise—that discretion should be explicit and auditable, not eliminated—is a direct engineering translation of due-process values [23]. The same logic recurs across the field: IRAC-aligned task decomposition in benchmarks and agent pipelines [3], factor-based retrieval descended from HYPO [21], and schema-constrained argument generation all re-impose the field's classical structures on top of statistical generation. Where Section 3 identified rule application as the unsolved component, the neuro-symbolic program treats that step as the proper domain of formal methods, with language models handling extraction and drafting around them. The scaffold, not the model, carries the normative weight.

4.3. LLM-as-Judge Under Structural Scrutiny

The third response concerns how outputs are evaluated at scale. LLM-as-a-judge evaluation—using a strong model to grade candidate outputs against criteria—has entered legal RAG assessment as a rapid alternative to human review; the LeMAJ framework applies legal-grounded rubrics to evaluate document recommendations end-to-end [24]. Its legitimacy, however, depends on the general judge-bias literature. Zheng et al. documented position bias, verbosity bias, and self-enhancement bias in LLM judges [25]; each has an unexamined legal analogue—a judge may favor longer recitations of precedent, defer to confidently cited authority, or reward fluency over citation accuracy—and a judge that hallucinates does not become reliable by grading others.

The bias experiments' design carries the lesson, because each bias was demonstrated by minimal paired perturbation: the same answer presented in both orders, the same content at two lengths, a model grading its own output against a competitor's. Small, controlled contrasts, large systematic effects—the signature of a measurement instrument with a causal structure of its own [25]. Transposed into legal evaluation, the perturbations write themselves, and their predicted effects are professionally specific: swap the order of two document recommendations and watch preference follow position; lengthen one memo's recitation of authorities and watch it defeat the terser, more accurate one; ask a model to grade legal-RAG outputs built on its own generations and measure the self-preference that would make a vendor's in-house evaluation quietly self-confirming [25]. None of these analogue studies has yet been run at scale in the legal domain, which is precisely the open problem: the general bias literature establishes that the instrument flexes, the legal LeMAJ-style frameworks show the instrument being deployed [24], and the field stands where general LLM evaluation stood before Zheng et al.—using judges at scale with an unfilled audit file on their biases [24,25]. The structural correctives already visible—fixed rubrics, reference answers, independent citation verification—are the legal translation of the general literature's mitigation suite; what converts them from recommendations into reliability is exactly the meta-evaluation this subsection exists to demand [24]. The emerging corrective is again structural: fixed rubrics, reference answers, and citation-verification steps that bind the judge to checkable criteria [24]. Meta-evaluation is where the field's measurement practices themselves become an accountability object—a fitting close to the CRA arc, because the framework's three dimensions here meet: capability is judged, reliability is at stake in the judging, and the rubric is the structure that makes the judgment auditable.

5. Conclusion

This survey organized the legal-LLM literature along a capability–reliability–accountability axis and reached three findings. First, capability measurement has matured across three generations—from task suites (LexGLUE [10]), to crowd-curated reasoning taxonomies (LegalBench [3]), to production-pipeline evaluation (BigLaw Bench [1])—with frontier models now exceeding 85% on aggregate legal-reasoning metrics. Second, a structural reliability gap separates those scores from deployed trustworthiness: benchmark validity limits [16], residual citation hallucination under grounding [17], and pipeline-level retrieval failures [1] are independent failure sources that aggregate metrics do not separate, which is why high scores and sanctioned briefs coexist. Third, accountability responses—high-risk regulation [19], neuro-symbolic scaffolding derived from the symbolic tradition [23], and rubric-bound judge evaluation [24]—converge on a single strategy: re-embedding statistical systems in inspectable structure.

The framework travels beyond law. Any domain combining generative models, high stakes, and licensed professionals—medicine and finance being the nearest cases—faces the same sequence: capability benchmarks arrive first, deployment exposes a reliability gap, and regulation plus structure-restoring architectures close it. For legal AI specifically, the most useful open problems follow the same axis: longitudinal benchmarks that resist saturation and isolate the rule–fact application step; end-to-end reliability metrics for RAG pipelines on professional corpora; and empirical study of judge-model bias in legal evaluation. The distance from benchmarks to courtrooms is real—but the field has begun to measure it, which is the first step toward closing it.

The framework's final yield is a reading of the field's near future that does not depend on prediction markets. On the capability axis, saturation is the visible destination: recall and issue-spotting tasks are near ceiling, and the marginal information in new aggregate scores is approaching zero—which pushes measurement toward the boundary tasks (application, situated judgment) and toward pipeline-level metrics, exactly where BigLaw Bench has already moved [1]. On the reliability axis, the citation-verification infrastructure is commoditizing: database-validated scoring of the kind the landmark audit introduced [17] is cheap, mechanical, and increasingly assumed, which will shift competitive differentiation toward the failure modes verification cannot catch—mischaracterized authority, retrieval omissions—and toward the pipeline diagnostics that expose them. On the accountability axis, the direction is legislative and slow: obligations are attaching to documentation, oversight, and logging faster than to

performance thresholds, because the first three are administrable and the last is not [19]. The CRA framework's prediction, stated plainly, is that these three currents converge on the same professional settlement law has always reached: competence assumed, diligence audited, liability assigned. The systems will get better; the structure of trusting them will not change—and the survey's contribution is to have shown that the structure, not the systems, is where the benchmark-to-courtroom gap actually lives.

References

[1] Harvey. (2025). BigLaw Bench: A deep dive on retrieval. Harvey AI. https://www.harvey.ai/blog/biglaw-bench-retrieval

[2] Mata v. Avianca, Inc., No. 22-cv-1461 (S.D.N.Y. 2023).

[3] Guha, N., Ho, D. E., & Nyarko, J. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning of large language models. In Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track).

[4] Hou, Z., Ye, Z., Zeng, N., Hao, T., & Zeng, K. (2025). Large language models meet legal artificial intelligence: A survey. arXiv preprint arXiv:2509.09969.

[5] Dehghani, F. (2025). Large language models in legal systems: A survey. Humanities and Social Sciences Communications, 12. https://www.nature.com/articles/s41599-025-05924-3

[6] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. In Findings of the Association for Computational Linguistics: EMNLP 2020.

[7] Tagarelli, A., & Simeri, A. (2023). Italian-legal-bert models for improving natural language processing in the Italian legal domain. Computer Law & Security Review, 48, 105787.

[8] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., & Androutsopoulos, I. (2019). Neural legal judgment prediction in English. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35.

[10] Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., & Aletras, N. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.

[11] Fei, Z., Shen, X., Zhu, D., Zhou, F., Han, Z., Zhang, S., Chen, K., Zhang, Z., Xu, E., Lin, B. Y., Geiger, A., Yu, T., & Chen, H. (2023). LawBench: Benchmarking legal knowledge of large language models. arXiv preprint arXiv:2309.16289.

[12] Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. B. (2023). ChatGPT goes to law school. Journal of Legal Education, 71(3), 387.

[13] Bommarito, M. J., II, Katz, D. M., & Detterman, A. (2022). GPT takes the bar exam. arXiv preprint arXiv:2112.01821.

[14] OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

[15] Hu, Y., Yu, Y., Gan, L., Wei, B., Kuang, K., & Wu, F. (2025). Evaluating test-time scaling LLMs for legal reasoning: OpenAI o1, DeepSeek-R1, and beyond. In Findings of the Association for Computational Linguistics: EMNLP 2025.

[16] Medvedeva, M., Wieling, M., & Vols, M. (2023). Rethinking the field of automatic prediction of court decisions. Artificial Intelligence and Law, 31(2), 195-212.
[17] Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. Journal of Legal Analysis, 16(1), 64-93.
[18] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33.
[19] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.
[20] Rissland, E. L., & Ashley, K. D. (1987). HYPO: A case-based system for trade secrets law. In Proceedings of the First International Conference on Artificial Intelligence and Law (ICAIL '87). ACM.
[21] Bruninghaus, L., & Ashley, K. D. (2003). Predicting outcomes of case-based legal arguments. In Proceedings of the 9th International Conference on Artificial Intelligence and Law (ICAIL '03). ACM.
[22] Bench-Capon, T. J. M. (2003). Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation, 13(3), 429-448.
[23] Chen, L., Cai, Y., Hou, Z., & Dong, J. S. (2025). Towards trustworthy legal AI through LLM agents and formal reasoning. arXiv preprint arXiv:2511.21033.
[24] Pradhan, A., Ortan, A., Verma, A., & Seshadri, M. (2025). LLM-as-a-judge: Rapid evaluation of legal document recommendation for retrieval-augmented generation. arXiv preprint arXiv:2509.12382.
[25] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems 36 (Datasets and Benchmarks Track).