Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FROM MODEL ACCURACY TO ACCOUNTABLE RELIANCE A Versioned Legal-Governance Framework for Generative-AI Reliance in Legal Services across the European Union, United Kingdom and United States

MUNAZA KHALID

munazzahkhalidedu@gmail.com

Southeast University https://orcid.org/0009-0006-1821-9782

Research Article

Keywords: generative AI, AI governance, legal services, accountable reliance, human oversight, professional responsibility, contestability, comparative regulation, versioning, CARI

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11823932/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

FROM MODEL ACCURACY TO ACCOUNTABLE RELIANCE

A Versioned Legal-Governance Framework for Generative-AI Reliance in Legal Services across the European Union, United Kingdom and United States

Evidentiary data cut-off: 1 August 2026

Abstract

Model accuracy alone cannot answer the core governance question raised by generative AI in legal practice: when may a lawyer, firm, or public body defensibly rely on a given AI configuration, on what evidence, and for how long? This study develops a versioned legal-governance protocol for the European Union, United Kingdom, and United States. Drawing on a preregistered 202-query legal benchmark, two randomized trials, 2026 professional and consumer surveys, provider documentation, and twelve comparative legal safeguards, the framework keeps four evidentiary domains distinct: task reliability (R_L), verified organizational governance (G), tested reviewer readiness (H), and jurisdictional legal context (J). The Constrained Accountable Reliance Indicator (CARI) applies a non-compensatory eligibility gate across R_L, G, and H, supported by three coupled algorithms (SHA-256 versioned gating, conformal selective abstention, and Page-Hinkley online drift expiry), exact statistical verification of empirical findings, and a preregistered four-phase local calibration protocol.

Keywords: generative AI; AI governance; legal services; accountable reliance; human oversight; professional responsibility; contestability; comparative regulation; versioning; CARI

Key contributions

- Frames legal-AI adoption as a problem of accountable reliance rather than model accuracy or product reputation alone.
- Separates jurisdictional legal obligations from three locally auditable conditions: task reliability, implemented governance and demonstrated reviewer readiness.
- Uses CARI as a prospective non-compensatory governance rule, not as a legal-compliance score, product ranking or safety certification.
- Treats meaningful human oversight as an empirically testable institutional control rather than a nominal human-in-the-loop requirement.
- Allocates evidence burdens across providers, deploying organizations, reviewers and affected persons, with explicit suspension, contestability and redress triggers.

- Makes reliance version-bound and expiring, so a material model, corpus, workflow or legal change reopens the governance decision.

1. Introduction

Generative artificial intelligence has rapidly transitioned from isolated predictive legal analytics into general-purpose and retrieval-augmented systems supporting legal research, contract drafting, document review, and summarization. Preregistered legal-research benchmarks demonstrate that even domain-specific retrieval-augmented tools remain materially fallible, combining accurate answers with persistent hallucinations and incomplete responses [1,2]. Experimental trials further show that while generative models deliver consistent completion-time savings, statistically supported quality improvements are task-specific and depend heavily on workflow design and structured user training [3–6]. At the same time, global professional and consumer surveys reveal accelerating organizational uptake (reaching 40% overall and 47% in corporate legal departments, while only 18% collect formal outcome metrics), alongside firm-size disparities, role-differentiated opacity, and strong public demand for enforceable protection and redress [7–10]. Technical studies of legal natural-language processing, retrieval pipelines, and frontier system cards confirm that legal factuality depends critically on retrieval architecture, query framing, and selective abstention rather than headline accuracy alone [11–13].

Yet rapid adoption has created an acute governance gap: institutional accountability lags behind everyday use while the underlying software keeps changing. Documented AI incidents climbed from 233 in 2024 to 362 in 2025, exposing lawyers, clients, and public bodies to unmeasured error risks in filings, advice, confidentiality, and access to remedies [14]. Across the European Union, United Kingdom, and United States, statutory and professional regimes (from the EU AI Act and UK regulatory, data, and judicial guidance to US ethics opinions, NIST risk frameworks, GAO standards, and FTC enforcement) set out duties of competence, supervision, documentation, human oversight, transparency, and contestability [15–24]. Formal law on the books, however, does not show whether a firm's controls actually work in daily practice. At the same time, frontier providers release frequent model and system-card updates (27 documented updates across OpenAI, Anthropic, and Google DeepMind over six months), creating silent performance drift that outdates static product-name approvals [25–27]. Prior work on law and AI, legal benchmarks, legal-aid pilots, and unmet legal need treats technical accuracy, human review, and legal

compliance in separate tracks, leaving organizations without a clear, auditable rule for authorizing, monitoring, or halting reliance on a specific AI setup [28–33].

This article addresses that gap through a versioned, empirically grounded, and executable framework for accountable reliance. Building on recent scholarship covering legal ethics, comparative AI governance, governance-by-design, corporate monitoring, and meaningful human oversight [34–40], we propose the Constrained Accountable Reliance Indicator (CARI). CARI keeps jurisdictional law (J) legally prior while demanding three non-substitutable local proofs: conservative task reliability (R_L), verified operational governance (G), and seeded-error-tested reviewer readiness (H). Rather than leaving prospective calibration as an open aspiration, we pair the governance rule with exact Wilson sample-size power curves, sequential Bayesian Beta-Binomial updating, a fatigue-adjusted human-oversight equation, Monte Carlo Dirichlet weight-sensitivity analysis, and three coupled AI algorithms. Four main contributions follow:

- Reframes legal-AI governance from abstract model accuracy to configuration-specific accountable reliance, separating jurisdictional law (J) from locally auditable R_L, G, and H.
- Operationalizes CARI via three coupled AI algorithms: SHA-256 versioned gating with missing-evidence blocking (Algorithm 1), conformal selective-prediction abstention routing (Algorithm 2), and Page-Hinkley/SPRT online drift expiry (Algorithm 3).
- Establishes exact statistical rigor across authentic secondary streams, including 81-set rounding-robust chi-square and Shannon entropy tests, Wilson sample-size power bounds, Bayesian R_L updating, fatigue-adjusted oversight modeling, and 2,000-run Dirichlet legal-matrix sensitivity.
- Resolves the prospective-validation bottleneck by specifying an executable, preregistered four-phase local calibration protocol linking actor-level evidence burdens, contestability logs, and automatic expiry triggers.

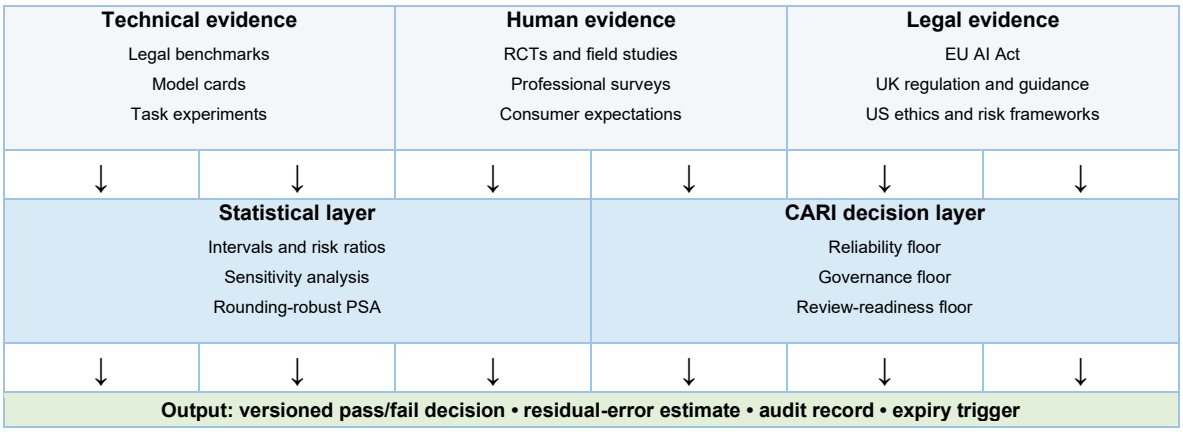


Fig. 1. Integrated evidence and decision architecture. Source: Author synthesis of the study design.

Figure 1 summarizes the evidentiary architecture: technical, human and legal sources retain different inferential roles until they are translated into a version-specific governance decision and audit record.

1.1 Research gap and legal-governance contribution

Table 1. Research gap and contribution map

Gap in existing work	Governance consequence	Contribution in this study
Legal-AI benchmarks and legal duties are usually analysed in separate literatures.	Benchmark accuracy can be mistaken for evidence that reliance is legally or professionally defensible.	Links version-specific reliability to implemented governance and demonstrated reviewer competence, while keeping legal context analytically separate.
Heterogeneous metrics are casually compared or averaged.	Macro-F1, open-ended accuracy, hallucination, incompleteness and survey attitudes lose their distinct legal significance.	Keeps metrics task-specific and failure-mode specific; only within-study contrasts and transparent sensitivity analyses are used.
Human-in-the-loop claims are often binary or role-based.	Nominal human presence can be mistaken for meaningful oversight even when reviewers cannot verify or intervene.	Requires a seeded-error reviewer test and evidence of authority, source access, escalation and correction; professional title alone cannot establish H. [39,40]
Threshold frameworks can appear objective without legal or empirical calibration.	Numerical floors risk being mistaken for safe harbours or standards of care.	Treats floors as ex ante governance parameters requiring prospective calibration; CARI is expressly not a compliance or liability test.

Gap in existing work	Governance consequence	Contribution in this study
Comparative AI regulation often stops at formal coverage.	Law-on-the-books can be mistaken for implementation, enforcement or organizational compliance.	Maps safeguards functionally as jurisdictional context and evidence sources; organization-level G still requires local proof of operation.
Recent work addresses legal ethics, governance-by-design, accountability and meaningful oversight, but usually as separate governance problems.	A deployment may satisfy general principles yet still lack the evidence needed to justify reliance on a specific task, workflow and version.	Combines R_L, operational G and tested H as non-substitutable local conditions, adds missing-evidence blocking, bottleneck attribution, contestability and expiry. [34-40]

Table 1 identifies the point at which existing benchmark, governance and oversight literatures leave a practical legal-governance gap, and states how the present framework responds without treating CARI as a substitute for law.

Five research questions organize the analysis. RQ1 asks how much uncertainty surrounds published legal-research accuracy and hallucination rates. RQ2 asks whether productivity evidence supports general quality claims or only task-specific gains. RQ3 asks how training, organizational adoption and consumer expectations affect the conditions for meaningful reliance. RQ4 asks how EU, UK and US legal architectures allocate safeguards, evidence obligations, contestability and professional responsibility. RQ5 asks what local evidence a legal organization should possess before it can justify, continue or suspend reliance on a particular AI configuration.

2. Related work: legal AI, oversight and governance

2.1 Legal-AI performance and hallucination

Magesh et al. reported the first preregistered evaluation of leading commercial legal-research tools across more than 200 queries. Retrieval grounding reduced, but did not eliminate, false or misleading responses. Dahl et al., using general-purpose models and specific verifiable questions about federal cases, observed considerably higher hallucination rates. The studies are not directly poolable because they examine different tasks and systems. Read together, however, they show that legal factuality depends on product architecture, question type, jurisdiction and evaluation rubric. [1,2]

Generic system-card evaluations provide useful context but cannot be transferred directly to legal work. OpenAI's SimpleQA results, for example, distinguish accuracy, hallucination and

abstention. That distinction matters because a system may improve by declining to answer rather than by producing more correct answers. SimpleQA does not test legal authority, jurisdictional validity, citation fidelity or omitted exceptions, so it is used here to clarify the structure of factuality rather than as a substitute for a legal benchmark. [13]

2.2 Human performance, training and workflow

Randomized evidence also shows why legal-AI performance cannot be reduced to access alone. The first randomized trial of GPT-4-assisted legal work found consistent time savings across four realistic tasks, but statistically supported quality improvement only for contract drafting. Subsequent studies suggest that workflow design matters: highlighting can reduce time without measurable quality loss, legal RAG and reasoning tools can improve quality-adjusted productivity on some tasks, and brief training can increase both adoption and examination performance. The evidence therefore supports neither the view that access automatically improves legal work nor the view that current limitations make the technology uniformly unproductive. [3-6] A field study of 91 legal-aid professionals provides complementary implementation evidence: 90% perceived some productivity improvement and 75% intended to continue using the tools. Because tool use and outcomes were largely self-reported, those results are not treated as an unbiased estimate of product effectiveness. [32]

2.3 Institutional trust, professional responsibility and governance

Trust within professional settings is distributed across roles. Developers and technical validators require diagnostic evidence about model behaviour; lawyers and other end users need accessible authorities, clear task boundaries and escalation rules; organizational approvers need a defensible basis for authorizing use. This division of labour can support efficient governance, but it can also create delegated opacity when users rely on institutional approval without knowing whether the system has moved outside its validated boundary. Professional reliance therefore requires more than confidence in a provider or a nominal review step: it requires an evidentiary chain linking technical validation, organizational controls, reviewer competence and accountable authorization. [10,34,39,40]

The three jurisdictions considered here distribute that accountability through different legal mechanisms. The EU AI Act establishes horizontal lifecycle duties for covered systems and classifies some justice-sector uses as high risk, without making every legal-service use a high-risk

system. The UK relies more heavily on distributed regulation, professional obligations, data safeguards and judicial guidance. The United States combines professional ethics, generally applicable law, agency enforcement and voluntary risk frameworks such as NIST and GAO. These architectures differ in legal form, but each raises the same operational question: what evidence shows that a particular organization has translated applicable duties into controls that actually function in the intended workflow? [15-24]

2.4 Positioning against recent law and AI-governance frameworks

Recent scholarship narrows the scope of any novelty claim. Terzidou examines competence, confidentiality, verification, and training in legal practice; Sourdin and Li compare organizational models of AI governance; Weinstein studies governance-by-design in public-facing legal guidance tools; Kourabas and Tsang assess heightened accountability in corporate monitoring; Laux shows that meaningful human oversight depends on institutional design rather than nominal sign-off; and Zhu et al. outline practical design patterns for oversight and contestability [34–40]. Taken as a whole, these studies confirm that governance, auditability, and retrieval grounding are already well-established principles.

The narrower problem taken up here is operational: how can a legal organization turn those general principles into a version-specific decision on whether reliance is currently justified? CARI answers that question by requiring three non-substitutable local proofs (R_L, operational G, and tested H), while keeping jurisdictional law J as the interpretive layer that sets applicable duties and safeguards. Adding a missing-evidence block, bottleneck attribution, an expiry rule, and a role-based audit log lets the framework supplement legal doctrine and professional rules without replacing them.

2.5 The governance problem: translating duties into auditable reliance

Legal and professional rules usually express duties at a high level of generality: competence, supervision, confidentiality, transparency, human oversight, documentation, and redress. A deployment decision, by contrast, calls for evidence about a concrete socio-technical setup. Law is therefore treated here neither as a numeric score nor as a footnote to technical testing. Jurisdictional law identifies the duties, actors, and remedies; local governance converts those duties into working controls; reviewer testing checks whether human oversight holds up in

practice; and task evaluation documents the configured system's error profile. Turning abstract rules into inspectable evidence is the core function of the framework [15–24,34–40].

2.6 Technical failure mechanisms of generative AI in legal workflows

A rigorous legal-governance framework cannot treat generative AI as an undifferentiated black box characterized only by headline accuracy. At the implementation level, ten specific technical and socio-technical mechanisms govern whether an AI configuration remains reliable in legal practice, and each maps directly onto CARI's local evidence requirements (R_L, G, and H):

(1) Decoding stochasticity and hardware non-determinism: Temperature ($T > 0$), nucleus (top-p) and top-k sampling, and GPU floating-point batching non-determinism mean that identical legal queries can yield divergent authorities or holdings across runs. Accountable reliance therefore requires logging exact decoding parameters and testing repeatability rather than relying on single-run outputs. (2) Retrieval-Augmented Generation (RAG) pipeline failure modes: Even grounded legal systems (see Figs. 16–17 in Appendix F) fail through chunking boundary splits (separating a statutory rule from its proviso or exception across text chunks), embedding drift and vocabulary mismatch (where dense vector search misses exact statutory citations unless paired with lexical BM25 hybrid retrieval), and vector-index staleness (retaining overruled precedent or superseded regulations). (3) Context-window degradation ('lost in the middle'): Long-context transformers exhibit degraded attention over mid-document clauses in lengthy contracts, discovery productions, and appellate records, creating systematic omission risks that short-query benchmarks do not expose.

(4) Sycophancy and leading-prompt sensitivity: Instruction-tuned models frequently defer to user presuppositions; a lawyer posing a leading legal query can induce confirmation-biased authority selection or fabricated support. (5) Confidence miscalibration vs. selective prediction: Linguistic fluency and authoritative tone do not track factual validity. Without empirical calibration (e.g., Expected Calibration Error) or selective-prediction abstention gates, overconfident errors bypass routine review. (6) Silent provider drift: Hosted API providers routinely apply quantization, distillation, safety-filter adjustments, and hidden system-prompt modifications without incrementing major model names, causing unannounced degradation in R_L. (7) Agentic compounding error and indirect prompt injection: Multi-step agentic legal workflows (already adopted by 15% of surveyed organizations; Fig. 22) suffer multiplicative reliability decay (if each

step succeeds with probability p, a k-step autonomous chain fails with probability 1 − p^k), while ingesting adversarial third-party filings creates indirect prompt-injection vulnerabilities.

(8) Benchmark contamination and data leakage: Public legal benchmarks may overlap with frontier pre-training corpora, inflating external R_L relative to novel client matters. (9) Confidentiality, privilege, and Zero-Data-Retention (ZDR) architecture: Public versus enterprise API tiers (Fig. 19) differ sharply in prompt logging, model-training opt-outs, and transient caching, directly implicating attorney-client privilege, legal professional secrecy, and data-protection obligations. (10) Human-AI cognitive failure modes: Nominal human-in-the-loop oversight degrades under automation bias (over-trusting fluent outputs), vigilance decrement (declining error detection as review volume rises), and progressive deskilling among junior practitioners who lose independent verification proficiency.

2.7 Critical literature synthesis and consolidated research gaps

Across the forty primary legal, empirical, and technical sources reviewed in Sections 2.1–2.6, research on generative AI in legal services falls into four largely separate tracks: (i) empirical legal-AI benchmarks and RAG engineering [1,2,11,12,31], which measure retrieval fidelity and hallucination rates without setting institutional authorization rules; (ii) randomized legal-work and training studies [3–6,10,32], which document task-specific speed gains and role-based opacity without turning reviewer skill into auditable gates; (iii) comparative AI regulation and legal-ethics studies [15–24,34–37], which map statutory and professional duties across the EU, UK, and US without showing how a firm proves live control execution; and (iv) human-oversight and algorithmic-assurance work [25–27,38–40], which critiques nominal human-in-the-loop claims without tying them to statistical power, selective-prediction routing, and online drift expiry.

Table 1b consolidates these four literature streams and pinpoints the exact operational gap addressed by this article: no existing framework integrates conservative task reliability (R_L), verified operational governance (G), fatigue-aware reviewer readiness (H), and jurisdictional legal context (J) into an executable, version-hashed algorithmic protocol (Algorithms 1–3) backed by exact statistical verification.

Table 1b. Literature Synthesis Matrix and Consolidated Research Gap

Literature Stream	Key Representative Studies	Core Established Finding	Unresolved Operational Gap Addressed Here
1. Legal-AI Benchmarks & RAG	[1,2,11,12,13,31]	RAG lowers but does not eliminate legal hallucination (17–33%); abstention differs from error.	No sample-size power rule, no local Bayesian updating, and no link to legal authorization.

Table 1b. Literature Synthesis Matrix and Consolidated Research Gap			
2. Human-AI Legal Work & Training	[3–6,10,32]	Time savings are broad (12–32%) but quality gains are task-specific; training doubles adoption odds.	No fatigue-adjusted oversight equation or seeded-error reviewer gate (H) for live workflows.
3. Comparative AI Law & Legal Ethics	[15–24,34–37]	EU (horizontal), UK (distributed), and US (ethics/agency) allocate duties differently.	Formal law-on-the-books is conflated with firm-level compliance; no weight-sensitivity test.
4. Oversight & Algorithmic Assurance	[14,25–27,38–40]	Nominal HITL becomes rubber-stamping; frontier models drift continuously.	No unified non-compensatory algorithm combining SHA-256 versioning, conformal abstention & SPRT expiry.
Source: Author synthesis of the 40-source evidentiary register (cut-off 1 August 2026).			

3. Data and methods

Figure 48 summarizes the five-stage empirical-legal research design that governs data ingestion, statistical verification, comparative legal coding, algorithmic formalization, and versioned audit output across this study.

Fig. 48. Five-Stage Empirical-Legal Research Design & Algorithmic Validation Pipeline			
Stage	**Methodological Layer**	**Authentic Inputs / Methods**	**Primary Output / Manuscript Anchor**
STAGE 1 Evidence Ingestion	**Targeted Multi-Stream Synthesis (Cut-off: 1 Aug 2026)**	202-query legal benchmark [1]; 2 RCTs [3,6]; TRI (n=1,514) & LSB (n=1,000) surveys [7,8]; 12 EU/UK/US instruments [15–24]; 27 provider updates [25–27].	Inferential-boundary map (Table 2; App. A–B) preventing invalid cross-study pooling.
STAGE 2 Statistical Core	**Exact Uncertainty & Hypothesis Testing**	All-set rounding enumeration (81 sets); Wilson CI & power inversion (Eq. 1); 4×3 χ^2 & Cramér V; Shannon entropy; Fisher OR; Spearman ρ; Poisson rate.	Rounding-robust intervals (Figs. 3, 15, 33–38, 40–41) & Table 8b verification.
STAGE 3 Comparative Law	**Functional Safeguard Coding & Robustness**	12-safeguard EU/UK/US descriptive matrix (Table L2); Friedman Q test; 2,000-run Monte Carlo Dirichlet weight perturbation.	Jurisdictional context layer J (Table 8, Figs. 10, 27, 42) strictly separated from local G.
STAGE 4 AI Algorithms	**Executable Governance & Oversight Control**	Algorithm 1 (SHA-256 CARI-Gate); Algorithm 2 (Conformal Abstention Router); Algorithm 3 (Page-Hinkley/SPRT Drift); Bayesian R_L & fatigue Eq. (4b).	Non-compensatory gate mechanics, bottleneck tornado & fatigue curves (Figs. 43–47).
STAGE 5 Governance Output	**Auditable, Contestable & Expiring Reliance**	Actor-level evidence allocation (Table 9); suspension & human-reconsideration triggers; configuration expiry rules.	Versioned pass/withhold/reject decision + full reproducibility trail (Table O1).
Fig. 48. End-to-end empirical-legal research design linking secondary evidence streams, exact statistical verification, comparative law, and executable AI governance algorithms. Source: Author study design.			

3.1 Evidence base and inclusion rules

The evidence synthesis is targeted and reproducible rather than exhaustive. Sources were included when they provided at least one of the following: an authoritative legal rule or official guidance; a transparent legal-AI benchmark; randomized or field evidence on legal work; a survey with a stated sampling frame; or qualitative evidence directly relevant to professional reliance. Vendor material was retained only when it documented market claims, product characteristics or adoption and was not treated as independent validation of accuracy.

All quantitative, statistical, and structural figures (Figures 1–49) are constructed as 100% native, editable Word objects with zero rasterized PNG images, accompanied by a complete OriginPro/Excel master workbook (OriginPro_Master_Data_47_Figures.xlsx) and reproduction scripts indexed in Appendix O; figure captions therefore report the evidentiary source directly.

The evidentiary data cut-off is 1 August 2026. A final pre-submission novelty check on 12 August 2026 was used only to position the contribution against already-published comparator literature [34-38]; it did not alter the quantitative evidence base, legal coding or provider-documentation counts.

Table 2. Evidence streams and inferential boundaries

Evidence stream	Unit/sample	Primary variables	Permitted inference
Legal-research benchmark	202 queries × four systems	Accurate, incomplete, hallucinated	Bounded product/task performance at the tested version
Legal-work RCT	60 law students; 240 task completions	Time and blind-graded quality	Causal task-level effects in a simulated setting
Training RCT	164 law students	Adoption, grade points, rule accuracy	Causal effect of brief training in one examination context
Professional survey	1,514 professionals; 27 countries	Adoption, frequency, use cases, ROI measurement	Reported organizational and personal practice
Consumer survey	1,000 adults in England and Wales	Expected better/worse, safeguards	Consumer expectations, not observed service outcomes
Qualitative case study	21 professionals in two firms	Role, understanding, opacity	Mechanism evidence, not prevalence

Evidence stream	Unit/sample	Primary variables	Permitted inference
Legal coding	12 safeguards × three jurisdictions	Present, partial/sectoral, or no general mechanism identified	Descriptive formal context only; not an organization-level governance or safety score
Legal-aid field study	91 professionals	Perceived productivity, continued-use intention	Implementation evidence; self-report and non-random tool use

Table 2 prevents inappropriate pooling by specifying what each evidence stream can and cannot support.

3.2 Benchmark reconstruction and interval estimation

The published results for the 202-query benchmark are reported as rounded percentages. To avoid treating an arbitrary integer reconstruction as if it were observed data, every count triple that sums to 202 and rounds to the published percentages was enumerated. Wilson intervals are then reported as envelopes across all admissible reconstructions. The conversion to counts is used only for interval estimation and descriptive comparison. [1]

$$CI_W = [\hat{p} + z^2/(2n) \pm z\sqrt{\{\hat{p}(1-\hat{p})/n + z^2/(4n^2)\}}] \,/\, [1 + z^2/n] \quad (1)$$

All four systems answered the same 202 queries, but paired query-level outcomes were not available. Hallucination comparisons are therefore limited to descriptive risk ratios and risk differences across the rounding-consistent reconstructions. Independent-samples confidence intervals are not reported because they would ignore within-query dependence and give a misleading impression of precision. [1]

The Wilson envelopes are descriptive uncertainty summaries conditional on the 202-query evaluation set and the binomial approximation. They should not be interpreted as confidence intervals for the universe of legal questions, other jurisdictions or future product versions.

To make local R_L estimation statistically actionable rather than ad hoc, organizations must determine the minimum evaluation sample size (n) required for a conservative 95% Wilson lower bound to clear a predeclared CARI floor r_s. Figure 33 maps exact minimum sample-size requirements across observed local accuracies for the candidate low (r_s = 0.60), medium (r_s = 0.75), and high (r_s = 0.90) tiers, demonstrating why small informal firm pilots (e.g., n = 20–50 queries) cannot statistically establish high-stakes reliability. Complementarily, Figure 35 plots the decay of the 95% Wilson confidence interval half-width from n = 50 to n = 2,000 queries,

contextualizing the ±6.5 percentage-point precision boundary of the n = 202 Magesh et al. benchmark.

Fig. 33. Minimum Local Audit Queries (n) for 95% Wilson Lower Bound to Clear CARI Floor

Observed Acc. ($\hat{p}$)	Low Tier ($r_s = 0.60$)	Medium Tier ($r_s = 0.75$)	High Tier ($r_s = 0.90$)
65%	n = 369	Infeasible ($\hat{p} \leq r_s$)	Infeasible ($\hat{p} \leq r_s$)
70%	n = 93	Infeasible ($\hat{p} \leq r_s$)	Infeasible ($\hat{p} \leq r_s$)
75%	n = 41	Infeasible ($\hat{p} \leq r_s$)	Infeasible ($\hat{p} \leq r_s$)
80%	n = 24	n = 289	Infeasible ($\hat{p} \leq r_s$)
85%	n = 15	n = 73	Infeasible ($\hat{p} \leq r_s$)
90%	n = 11	n = 33	Infeasible ($\hat{p} \leq r_s$)
95%	n = 10	n = 19	n = 139
98%	n = 10	n = 14	n = 55

Green: n ≤ 100 (feasible small audit); Amber: 101–300 (comparable to Magesh n=202); Red: n > 300 or infeasible. Exact Wilson score inversion (Eq. 1).

Fig. 33. Minimum local audit sample size (n) required for the 95% Wilson lower bound to clear candidate CARI reliability floors (r_s = 0.60, 0.75, 0.90). Source: Exact Wilson score inversion (Eq. 1) using Table M2 candidate floors.

Fig. 35. 95% Wilson Confidence Interval Half-Width Decay Across Sample Sizes

Sample Size (n)	p = 0.50 (Max Var)	p = 0.65 (Lexis+)	p = 0.80 (Med Floor)	p = 0.90 (High Floor)
n = 50	±13.4 pp	±12.8 pp	±10.9 pp	±8.5 pp
n = 100	±9.6 pp	±9.2 pp	±7.8 pp	±6.0 pp
n = 202 (Benchmark)	±6.8 pp	±6.5 pp	±5.5 pp	±4.2 pp
n = 500	±4.4 pp	±4.2 pp	±3.5 pp	±2.6 pp
n = 1000	±3.1 pp	±3.0 pp	±2.5 pp	±1.9 pp
n = 2000	±2.2 pp	±2.1 pp	±1.8 pp	±1.3 pp

Exact 95% Wilson score half-widths (± percentage points). Highlights why n = 202 leaves ±6.5 pp uncertainty at p = 0.65.

Fig. 35. Statistical precision decay of the 95% Wilson confidence interval half-width across legal-benchmark sample sizes (n = 50 to 2,000), highlighting the n = 202 benchmark scale. Source: Exact Wilson score calculation (Eq. 1).

3.3 Functional comparative-law coding of governance safeguards

The comparative legal analysis uses a functional coding approach. Twelve safeguards were traced across the three jurisdictions: cross-sector binding AI law; justice-sector classification; ex ante risk management; data governance; technical documentation; logging; human oversight; accuracy, robustness and cybersecurity; transparency; contestability and redress; professional ethics guidance; and post-deployment monitoring or enforcement. Each is described as Present, Partial/sectoral, or No general mechanism identified. The categories are descriptive legal mappings, not measures of implementation or institutional quality. No aggregate jurisdictional

score is used in the primary analysis, and the legal matrix is not a proxy for organization-level G. [15-24]

3.4 CARI as an operational governance rule, not a compliance test

For consequential legal uses, a compensatory average can conceal a governance failure: strong technical performance cannot cure the absence of effective supervision, and extensive policies cannot cure an unreliable system. CARI therefore uses a non-compensatory eligibility rule for deployment x at stakes tier s. It is an internal assurance decision rule, not a statement that a deployment complies with all applicable law, discharges professional responsibility, or determines civil, disciplinary or regulatory liability:

$$CARI_s(x)=1\{R_L(x)\geq r_s \wedge G(x)\geq g_s \wedge H(x)\geq h_s\} \quad (2)$$

R_L denotes a conservative lower-bound estimate of material reliability for the defined task, frozen model, retrieval corpus and workflow. G measures whether the organization's applicable controls are demonstrably operating in practice. H captures reviewer competence in detecting and appropriately escalating seeded material errors under realistic workflow conditions. Jurisdictional safeguards provide context for these measurements but are not themselves G. If R_L, G or H has not been measured locally for the intended use, CARI eligibility for live deployment cannot be established.

$$AM_s(x)=\min\{R_L(x)-r_s,\ G(x)-g_s,\ H(x)-h_s\} \quad (3)$$

The assurance margin AM_s is the signed distance from the weakest predeclared floor and identifies the binding measured bottleneck. The thresholds r_s, g_s and h_s are governance parameters to be fixed before evaluation for a defined use and then calibrated prospectively against material errors, corrections and incidents. They are not statutory standards, safe harbours or product certification values. Jurisdictional context J remains legally prior: it determines which duties, constraints, disclosures, rights and remedies must be considered before the organization specifies its operational controls.

Table 3. Operational definitions and minimum evidence for accountable-reliance inputs

Component	Governance meaning	Minimum evidence	What must not be substituted
R_L: task reliability	Conservative lower bound on version- and task-specific material reliability; error types remain explicit rather than collapsed when consequences differ.	Representative local evaluation set; frozen model/corpus/prompt workflow; stated material-error taxonomy; source-verification protocol; interval estimate.	Generic factuality benchmarks, vendor claims, adoption surveys or unrelated legal tasks.
G: operational governance	Applicable legal, professional and organizational controls translated into measures that are verified as implemented and functioning for the intended workflow.	Access/data controls; confidentiality and professional-use rules where applicable; version/change management; logging; evaluation ownership; incident/escalation process; redress/complaint route; auditable evidence of operation.	Jurisdictional law-on-the-books, a written policy alone, or provider documentation.
H: reviewer readiness	Conservative lower-bound performance on detecting and appropriately escalating seeded material errors under realistic time and source-access conditions.	Predefined seeded-error test; authority verification; escalation accuracy; correction latency; repeat testing after material workflow change.	Professional title, training attendance, experience claims or a nominal human-in-the-loop step.
J: jurisdictional legal context (not a CARI score)	Interpretive legal layer identifying applicable duties, actors, safeguards, disclosure expectations, contestability and remedies for the use case.	Authoritative legislation, professional rules, regulator/judicial guidance and enforcement mechanisms.	An organization-level G score, a comparative safety ranking, a legal opinion, or evidence of actual compliance.

Table 3 defines R_L, G, and H as locally auditable inputs and keeps them strictly separate from jurisdictional legal context.

3.5 Sensitivity and prospective validation protocol

The available secondary sources do not report paired local measurements of G and H for the named legal-AI products. Product-specific CARI results therefore cannot be treated as empirical deployment approvals. For transparency, Figures 11-12 and Appendix M retain the earlier counterfactual stress test, but the product labels are suppressed in the probability display. The four

anonymized profiles preserve only benchmark-accuracy inputs; G, H and candidate floors remain declared assumptions, and benchmark accuracy is not itself a validated local R_L measure. The simulations illustrate gate mechanics and sensitivity only; they are not safety probabilities or product rankings.

A future probabilistic implementation should calibrate its distributions from repeated local evaluations, governance audits and reviewer tests. Until such data exist, simulated pass probabilities and bottleneck frequencies are treated as sensitivity diagnostics. Their purpose is to expose which assumptions drive a decision and to specify what would need empirical validation, not to supply a safe harbour.

To bridge the gap between external benchmarks and prospective local validation, CARI supports sequential Bayesian Beta-Binomial updating of local task reliability R_L. Rather than discarding external evidence or treating it as sufficient, a downweighted benchmark prior (e.g., Lexis+ 65% at effective sample size n_eff = 50) is updated as the deploying organization accumulates verified local audit outcomes. As illustrated in Figure 45, even if local evaluation observes 82% accuracy, a small n = 50 audit leaves the 5% Bayesian credible lower bound at 65.9% (below a 75% medium-tier floor), whereas scaling local verification to n = 500 tightens the posterior and lifts the 5% lower bound to 77.6%, providing a transparent, mathematically auditable stopping rule for prospective calibration.

Fig. 45. Bayesian Beta-Binomial Posterior Updating for Local R_L Assurance

Audit Stage	Post. Mean	90% Credible Int.	5% Lower Bound	Medium-Tier Gate (r_s = 75%)
Benchmark Prior (n_eff = 50, μ = 65%)	64.9%	**[53.6% – 75.5%]**	**53.6%**	**BELOW r_s = 75% (Withhold)**
After Local Audit n = 50 (82% obs.)	73.4%	**[65.9% – 80.3%]**	**65.9%**	**BELOW r_s = 75% (Withhold)**
After Local Audit n = 200 (82% obs.)	78.5%	**[74.2% – 82.7%]**	**74.2%**	**BELOW r_s = 75% (Withhold)**
After Local Audit n = 500 (82% obs.)	80.4%	**[77.6% – 83.1%]**	**77.6%**	**CLEARS r_s = 75%**

Downweighted Lexis+ prior (α_0=32.75, β_0=17.75) updated sequentially with local 82% accuracy audits. Only n=500 clears the 75% conservative floor.

Fig. 45. Bayesian Beta-Binomial posterior updating of local task reliability R_L from a downweighted benchmark prior (n_eff = 50, μ = 65%) across sequential local audits (n = 50, 200, 500 at 82% local accuracy). Source: Author Bayesian calibration model derived from Magesh et al. (2025) baseline.

3.6 Residual-error illustration

Human review can be separated into review coverage h and the conditional probability d that a covered material error is detected. For N outputs and an illustrative baseline material-error probability p_e, the expected number of residual errors, under the simplifying assumption that covered errors are detected independently with probability d, is:

$$E_res = N\,p_e\,(1 - h\,d) \quad (4)$$

The derivatives ∂E_res/∂h=−Np_ed and ∂E_res/∂d=−Np_eh make the interaction between coverage and detection explicit: the value of improving either control depends on the level of the other. Equation (4) is used only as an explanatory control model. It is not a malpractice forecast, and it assumes a common baseline error probability and a constant conditional detection rate. Benchmark hallucination rates should not be interpreted as live-deployment material-error rates without local validation.

Because empirical human-factors research demonstrates that constant detection competence d is unrealistic under high-volume review, where automation bias and vigilance decrement cause error detection to decay as review load increases, we extend Equation (4) to a fatigue-adjusted oversight model: E_res(h) = N p_e [1 − h · d_0 · exp(−λh)] (4b), where d_0 denotes baseline detection competence under low fatigue and λ ≥ 0 captures the rate of vigilance decrement and rubber-stamping as review coverage h approaches 100%. This non-linear formulation (visualized in Section 5.2 via Figures 43–44 alongside Appendix N Figures 29–31) directly explains why mandating 100% nominal human review without workload limits or interface support can yield diminishing governance returns.

3.7 Formal AI algorithmic architecture for accountable reliance

To turn the CARI governance rule into an executable AI workflow, we specify three coupled algorithms: (1) a version-hashed non-compensatory evaluation and bottleneck-attribution routine (Algorithm 1), (2) a calibrated selective-prediction and conformal abstention router (Algorithm 2), and (3) a sequential Page-Hinkley / SPRT online drift detector for silent provider updates (Algorithm 3). Used in sequence, these three routines link statistical learning controls directly to auditable legal governance.

Algorithm 1: Versioned CARI-Gate & Bottleneck Attribution (CARI_EVALUATE)

```
Input : Config C = (model_id, T, top_p, corpus_hash, prompt_hash, juris_J, stakes_s);
        Predeclared floors (r_s, g_s, h_s); Local audit datasets (D_rel, D_gov, D_rev)
Output: Decision in {APPROVE_VERSIONED, WITHHOLD_MISSING, REJECT_FLOOR}, AM_s,
Bottleneck, AuditHash
1.  config_hash <- SHA256(C)                          // Freeze exact socio-technical
configuration
2.  if (D_rel is empty) or (D_gov is empty) or (D_rev is empty) then
3.      return (WITHHOLD_MISSING, NULL, 'Missing Local Evidence', config_hash)
4.  R_L <- WilsonLowerBound95(correct(D_rel), |D_rel|) // Conservative task reliability
(Eq. 1)
5.  G   <- AuditControlCompliance(D_gov, juris_J)      // Verified operational controls
6.  H   <- WilsonLowerBound95(detected(D_rev), |D_rev|)// Seeded-error reviewer detection
7.  margins <- {'R_L': R_L - r_s, 'G': G - g_s, 'H': H - h_s}
8.  AM_s <- min(margins.values())                      // Assurance margin (Eq. 3)
9.  bottleneck <- argmin(margins)                      // Binding constraint attribution
10. if AM_s >= 0 then
11.     return (APPROVE_VERSIONED, AM_s, bottleneck, config_hash)
12. else
13.     return (REJECT_FLOOR, AM_s, bottleneck, config_hash)
```

Algorithm 2: Selective-Prediction & Conformal Abstention Router (SELECTIVE_ROUTE)

```
Input : Legal query q; Frozen RAG-LLM M_C; Calibrated confidence estimator s(q, y) in
[0,1];
        Conformal abstention threshold tau_alpha calibrated on local validation set D_cal
Output: Routed response y_out in {VERIFIED_DRAFT, ABSTAIN_ESCALATE} + AuditLog
1.  (y, citations, logprobs) <- M_C.generate(q)        // Generate answer with retrieval
trace
2.  c_score <- s(q, y, citations, logprobs)            // Compute calibrated confidence
score
3.  if (c_score < tau_alpha) or (citations.unverified > 0) then
4.      Log(config_hash, q, 'ABSTAIN_ESCALATE', c_score)
5.      return (ABSTAIN_ESCALATE, 'Insufficient grounding: routed to supervising lawyer')
6.  else
7.      Log(config_hash, q, 'VERIFIED_DRAFT', c_score)
8.      return (VERIFIED_DRAFT, y, citations)
```

Algorithm 3: Sequential Online Drift & Expiry Detector (SPRT_DRIFT_MONITOR)

```
Input : Stream of periodic seeded-error probe outcomes e_t in {0,1} (1 = material error);
        Approved baseline error rate p_0 = 1 - r_s; Drift tolerance delta > 0; Alarm
threshold gamma
Output: Status in {VALID_WITHIN_SPEC, EXPIRE_AND_REVALIDATE}
1.  Initialize cumulative drift statistic U_0 <- 0, m_0 <- 0
2.  for each audit probe t = 1, 2, ... do
3.      U_t <- U_{t-1} + (e_t - p_0 - delta)           // Page-Hinkley cumulative
deviation
4.      m_t <- min(m_{t-1}, U_t)
5.      if (U_t - m_t) > gamma or ProviderCardChanged(C) then
6.          TriggerSuspension(config_hash, t)          // Auto-expire versioned approval
7.          return EXPIRE_AND_REVALIDATE
8.  return VALID_WITHIN_SPEC
```

4. Results

4.1 Legal-research systems remain materially fallible

Among the four systems evaluated by Magesh et al., Lexis+ AI had the highest reported accuracy at 65%, followed by GPT-4 at 49%, Westlaw AI-assisted research at 42% and Ask Practical Law AI at 20%. These figures should not be read as a simple product ranking. Ask Practical Law's 63% incompleteness reflects a narrower corpus and a different refusal pattern, whereas GPT-4's 43% hallucination rate reflects greater responsiveness without comparable legal grounding. [1]

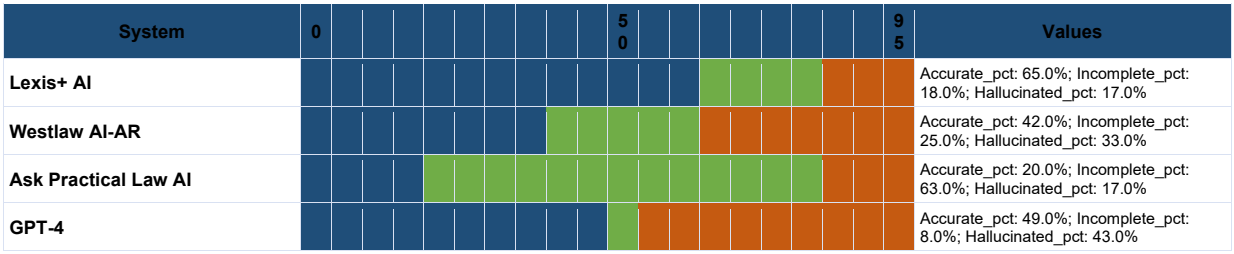


Fig. 2. Accurate, incomplete and hallucinated responses. Source: Reconstructed from Magesh et al. (2025), n=202 queries per system.

The benchmark composition in Figure 2 makes the central reliability problem visible: every tested system combines correct answers with a non-trivial share of incomplete or hallucinated responses, so accuracy alone is an incomplete description of performance.

The reconstructed Wilson envelopes remain wide enough to warrant caution but do not remove the observed differences. Approximate accuracy intervals are 58-72% for Lexis+ AI, 35-49% for Westlaw, 15-26% for Ask Practical Law and 42-56% for GPT-4. These intervals reflect query-sample uncertainty together with ambiguity introduced by the published rounding. They do not capture coding disagreement, benchmark-selection effects or product-version drift. [1]

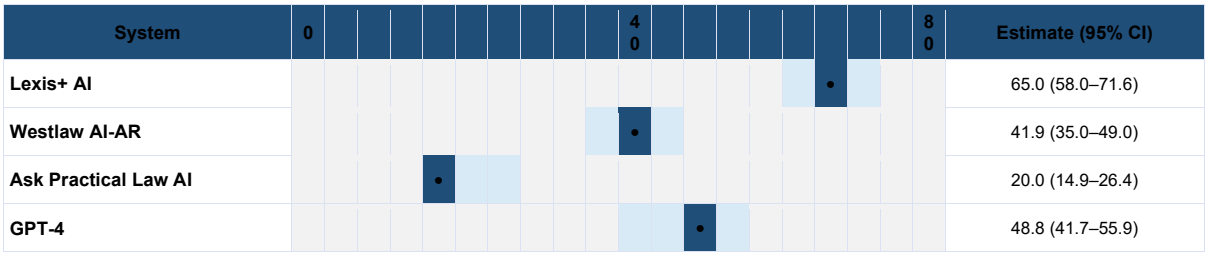


Fig. 3. Wilson intervals for legal-research accuracy. Source: Counts reconstructed from rounded percentages in Magesh et al. (2025).

The interval plot in Figure 3 emphasizes the uncertainty around reconstructed legal-research accuracy and avoids treating rounded published percentages as exact estimates.

Table 4. Legal-research outcomes and reconstructed Wilson intervals

System	Accurate	Incomplete	Hallucinated	Accuracy 95% CI	Hallucination 95% CI
Lexis+ AI	65%	18%	17%	58.0–71.6%	12.3–23.1%
Westlaw AI-AR	42%	25%	33%	35.0–49.0%	26.6–39.9%
Ask Practical Law AI	20%	63%	17%	14.9–26.4%	12.3–23.1%
GPT-4	49%	8%	43%	41.7–55.9%	36.0–50.0%

Table 4 reports the benchmark outcomes together with rounding-aware uncertainty, providing the quantitative basis for the legal-research reliability discussion.

The outcome pattern also illustrates why accuracy is not sufficient on its own. Westlaw and GPT-4 are relatively responsive but show substantial hallucination, whereas Ask Practical Law avoids some false answers by returning many more incomplete ones. Those failure modes call for different controls. An incomplete answer can be routed to escalation if the system makes the limitation visible; an apparently complete but false answer may pass unnoticed.

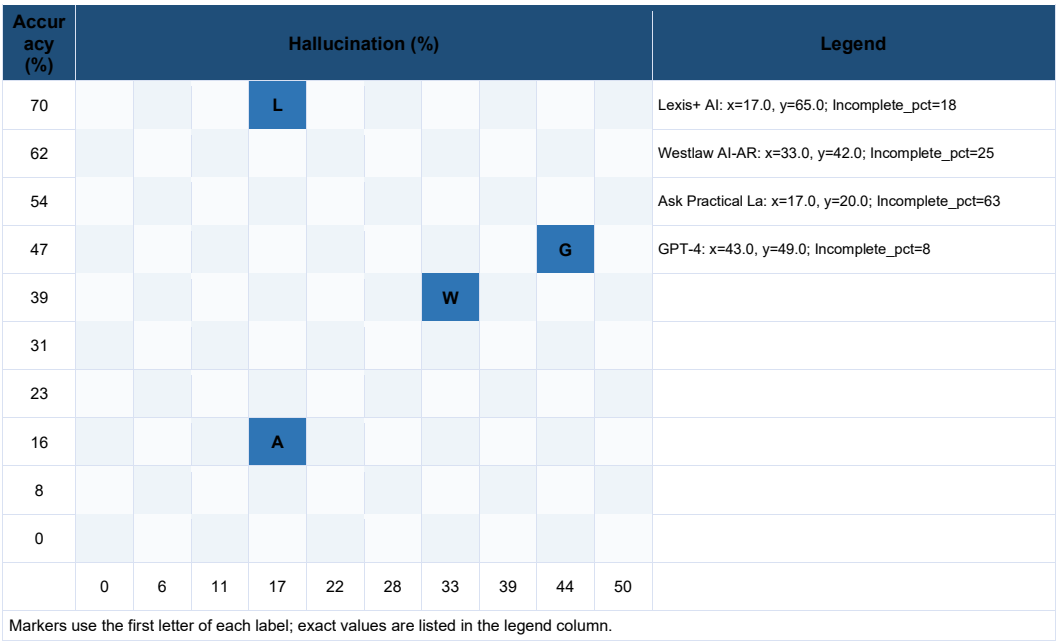


Fig. 4. Accuracy, hallucination and incompleteness trade-off. Source: Magesh et al. (2025).

By plotting accuracy against hallucination while retaining incompleteness as a separate dimension, Figure 4 shows why distinct failure modes call for distinct controls.

Figure 34 adds a full pairwise risk-difference forest plot (in percentage points, using exact rounding envelopes from Table C1) across all three outcome modes relative to Lexis+ AI, alongside the hallucination risk ratios in Appendix D (Figure 15) and the trade-off plot in Figure 4. Westlaw AI-AR shows a +15.8 pp higher hallucination rate and a −23.3 pp accuracy deficit, whereas Ask Practical Law AI matches Lexis+ AI on hallucination (+0.0 pp) while incurring a +45.0 pp incompleteness penalty. These real benchmark contrasts confirm that reliability governance must track each failure mode separately instead of collapsing them into a single accuracy figure.

Fig. 34. Rounding-Robust Pairwise Risk Differences vs. Lexis+ AI (Percentage Points)

System vs. Lexis+ AI	Accuracy Δ (pp)	Incompleteness Δ (pp)	Hallucination Δ (pp)
Westlaw AI-AR	**−23.3 pp**	+6.9 pp	**+15.8 pp**
Ask Practical Law AI	**−45.0 pp**	**+45.0 pp**	+0.0 pp
GPT-4	**−16.3 pp**	−9.9 pp	**+25.7 pp**

Computed across all 9 pairwise admissible integer reconstructions from Table C1 (n = 202 queries per system). Bar length proportional to |Δ| (scale 0–50 pp).

Fig. 34. Rounding-robust pairwise risk differences (percentage points vs. Lexis+ AI) across accurate, incomplete, and hallucinated response modes. Source: Exact enumeration over all admissible integer count sets in Table C1 (Magesh et al., 2025, n = 202).

4.2 Generic factuality confirms that abstention is a distinct outcome

SimpleQA reinforces the distinction between hallucination and abstention. In OpenAI's reported results, GPT-5 thinking achieved 55% accuracy and 40% hallucination, whereas GPT-5 thinking mini had lower accuracy but markedly lower hallucination, implying more abstention or other unsuccessful non-hallucinatory outcomes. These figures do not estimate legal accuracy. Their relevance here is narrower: in consequential work, a visible refusal that triggers escalation may be safer than a plausible but unsupported answer. [13]

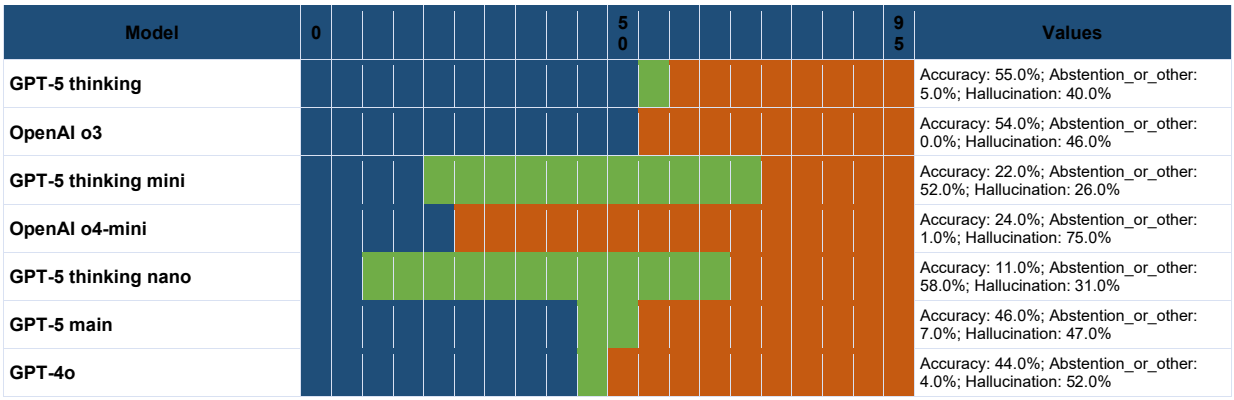


Fig. 5. Accuracy, hallucination and abstention on SimpleQA. Source: OpenAI GPT-5 System Card (2025); this is contextual factuality evidence, not a legal-task benchmark. [13]

Figure 5 is used only to illustrate that abstention is analytically distinct from hallucination; it motivates escalation and refusal design but is not transferred into a legal reliability estimate.

Quantifying the governance value of abstention requires separating raw accuracy from selective accuracy, defined as Accuracy / (1 − Abstention) (correctness conditional on the model choosing to answer). Figure 36 maps the selective-prediction efficiency frontier across the seven SimpleQA models in Figure 5 and Table E1. Models positioned well above the zero-abstention identity line (such as GPT-5 thinking mini, whose selective accuracy reaches 45.8% despite 22.0% raw accuracy due to 52% abstention) demonstrate how explicit refusal mechanisms concentrate errors into visible escalations rather than silent hallucinations.

Fig. 36. Selective-Prediction Efficiency Frontier on SimpleQA (Table E1)

Model	Raw Accuracy (%)	Abstention	Selective Acc. (%)	Selective Lift
GPT-5 thinking	55.0%	5.0%	57.9%	+2.9 pp
OpenAI o3	54.0%	0.0%	54.0%	+0.0 pp
GPT-5 main	46.0%	7.0%	49.5%	+3.5 pp
GPT-4o	44.0%	4.0%	45.8%	+1.8 pp
GPT-5 thinking mini	22.0%	**52.0%**	**45.8%**	**+23.8 pp**
GPT-5 thinking nano	11.0%	**58.0%**	**26.2%**	**+15.2 pp**
OpenAI o4-mini	24.0%	1.0%	24.2%	+0.2 pp

Selective Accuracy = Accuracy / (1 − Abstention). High-abstention models (highlighted) convert silent hallucinations into visible escalations.

Fig. 36. Selective-prediction efficiency frontier on SimpleQA comparing raw accuracy against selective accuracy on answered queries, shaded by abstention rate. Source: Author calculation from OpenAI GPT-5 System Card (2025, Table E1); contextual factuality structure only.

4.3 Productivity gains are substantial but quality gains are task-specific

In the Choi et al. trial, mean task-completion time fell by 11.8-32.1% across the four tasks. A descriptive aggregation of the published task means corresponds to a 19.0% reduction, although

this is not a participant-level pooled estimate. Contract drafting showed the largest time reduction and the client memorandum the smallest. [3]

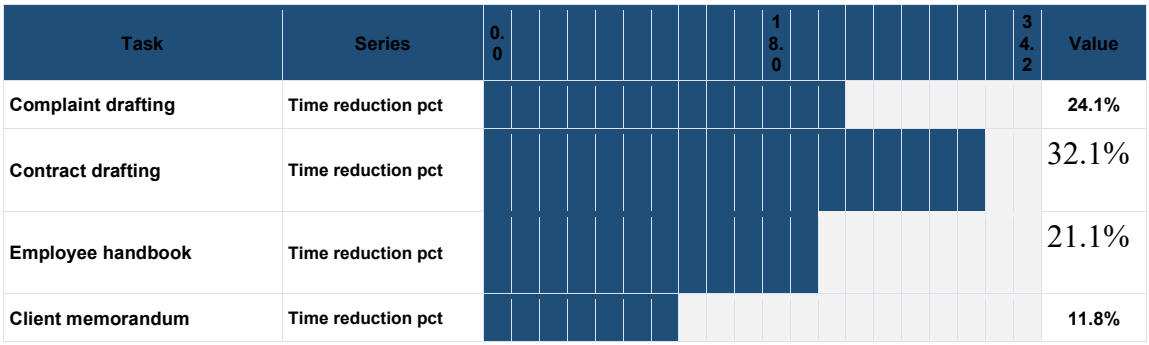


Fig. 6. Task-level reduction in mean completion time. Source: Choi, Monahan and Schwarcz (2024).

Across the four randomized tasks, Figure 6 shows consistent time savings but substantial variation in their magnitude.

Quality estimates were positive for three tasks, but only the contract-drafting effect was statistically supported at conventional levels; the client-memorandum estimate was slightly negative. The trial therefore does not support a general claim that access to a general-purpose model improves legal quality. It supports the more limited conclusion that effects vary by task and should be tested locally. [3]

Table 5. Task-level effects in the GPT-4 legal-work randomized trial

Task	No-AI minutes	AI minutes	Time reduction	Quality change	p-value
Complaint drafting	160.69	122.00	24.1%	+0.17	0.0862
Contract drafting	69.72	47.59	32.1%	+0.24	0.0060
Employee handbook	37.24	29.41	21.1%	+0.07	0.3532
Client memorandum	244.41	215.69	11.8%	-0.07	0.5980

Table 5 shows that productivity gains were broad but quality gains were task-specific, which argues against a single generalized productivity or quality claim.

Figure 37 places all four randomized tasks from Choi et al. (2024) on a single time–quality quadrant, combining the speed gains in Figure 6 with the quality estimates in Appendix G (Figure 18). Contract drafting is the only task in the statistically supported 'faster and better' region (+32.1% speed, +0.24 quality, p = 0.006). Client memorandum drafting, the most open-ended

doctrinal task, posts the smallest time saving (+11.8%) and a non-significant negative quality point estimate (−0.07, p = 0.598). These experimental results support CARI's rule that reliance must be approved per legal task rather than per software product.

Fig. 37. Time–Quality Trade-off Quadrant in Randomized Legal Work (Choi et al., 2024)

Legal Task	Time Reduction (%)	Quality Δ	p-value	Quadrant Classification
Contract drafting	−32.1%	**+0.24**	**p = 0.006***	**Faster & Statistically Better**
Complaint drafting	−24.1%	+0.17	p = 0.086 (ns)	Faster; Quality Uncertain
Employee handbook	−21.1%	+0.07	p = 0.353 (ns)	Faster; Quality Uncertain
Client memorandum	−11.8%	**−0.07**	p = 0.598 (ns)	Faster; Unimproved Quality

Mean task time reduction = 19.0%. Only contract drafting clears p < 0.05 for quality improvement (n = 240 completions).

Fig. 37. Time–quality trade-off quadrant in randomized GPT-4-assisted legal work, encoding task-level statistical significance. Source: Author synthesis of Choi, Monahan and Schwarcz (2024, Table 5, n = 240 completions).

4.4 Training is part of the intervention

Chen and Bao found that a brief training intervention increased LLM adoption from 26% to 41% (p=0.044) and improved examination performance by 0.27 grade points relative to untrained access (p=0.027). Untrained access did not reliably improve performance and was associated with shorter answers and more case misstatements. For the present framework, the implication is straightforward: H cannot be inferred merely from the presence of a human reviewer. Reviewer readiness must be demonstrated on the intended workflow. [6]

Condition	Outcome	Value
Optional LLM, no training	Adoption	26.00
Optional LLM + training	Adoption	41.00
Training vs untrained	Grade-point effect	0.27

Fig. 7. Training intervention effects. Source: Chen and Bao (2026), randomized study, n=164; the source is a preprint and is interpreted accordingly. [6]

The training contrast in Figure 7 supports treating reviewer preparation as part of the intervention rather than assuming that a nominal human-in-the-loop step is sufficient.

Figure 38 quantifies the statistical precision of the training intervention summarized in Figure 7 and Appendix H (Table H1), displaying both the +15 percentage-point adoption lift (26% to 41%, p = 0.044) with Wilson uncertainty bounds and the +0.27 grade-point substantive gain (95% CI: 0.03–0.51, p = 0.027). Because trained reviewers also improved on applicable-rule accuracy (p = 0.014), reviewer preparation directly shifts the empirical distribution of H and cannot be treated as an administrative afterthought.

Fig. 38. Causal Effect of Brief Training on Legal-AI Adoption & Exam Performance (n = 164)				
Outcome Measure	**Point Est.**	**95% CI**	**Interval Visual (● = Est.)**	**Statistical Interpretation**
LLM Adoption (Untrained)	26.0%	[17.5% – 36.1%]		Baseline arm (n ≈ 82)
LLM Adoption (Trained)	**41.0%**	[31.4% – 52.3%]		+15 pp lift (p = 0.044; OR = 2.06)
Exam Grade Gain (Trained vs. Untrained)	**+0.27 pts**	[+0.03 – +0.51]		p = 0.027; Rule accuracy p = 0.014
Source: Reconstructed from Chen & Bao (2026) randomized study (n = 164). Demonstrates training shifts reviewer readiness H.				

Fig. 38. Causal effect of brief legal-AI training on adoption rates (Panel A, Wilson 95% CIs) and examination grade-point performance (Panel B, approximate 95% CI). Source: Reconstructed from Chen and Bao (2026, randomized study, n = 164).

4.5 Adoption is moving faster than organizational measurement

Professional adoption is rising faster than many organizations are measuring its effects. In the 2026 Thomson Reuters survey, organizational use reached 40%, compared with 22% in the previous survey. Reported use was 47% in corporate legal departments, 41% in law firms and 32% in government legal organizations, while corporate risk respondents reported 62%. Only 18% of respondents said their organizations collected ROI metrics, and the metrics used were more often internal measures than indicators of client outcomes or error. [7]

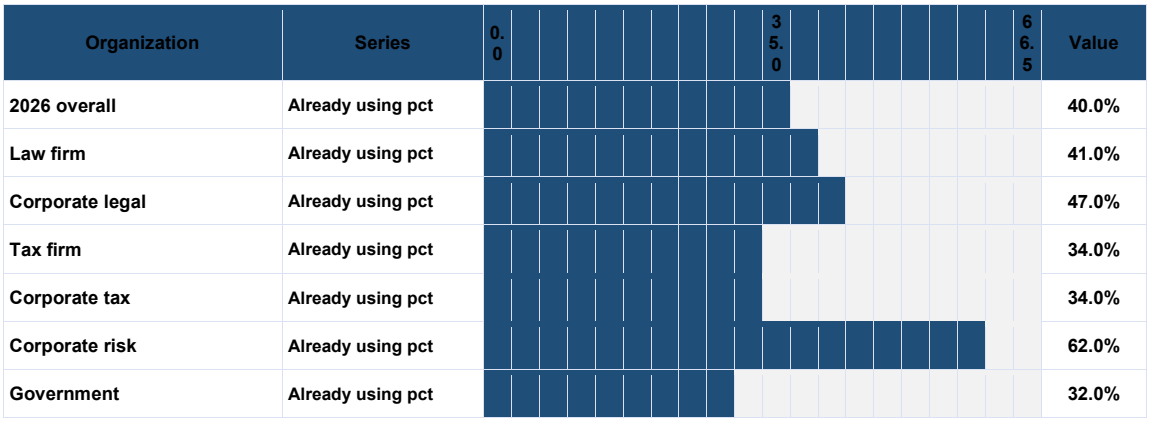


Fig. 8. Organizational GenAI use by sector. Source: Thomson Reuters Institute (2026), survey of 1,514 AI-familiar professionals in 27 countries. [7]

Figure 8 documents broad organizational adoption across professional-service sectors, reinforcing the gap between deployment exposure and formal outcome measurement.

Table 6. Professional adoption and measurement indicators

Indicator	Reported value	Interpretation
Organizations already using GenAI	40%	Up from 22% in the prior survey

Indicator	Reported value	Interpretation
Current users using GenAI at least weekly	82%	Use is embedded in regular work
Organizations collecting AI ROI metrics	18%	Measurement substantially trails adoption
Organizations already using agentic AI	15%	Early adoption with 53% planning or considering
Expect agentic AI central by 2030	77%	Anticipated workflow dependence
Corporate respondents wanting outside firms to use AI	Approximately two-thirds	Demand is not matched by clear mandates

Table 6 contrasts rapid organizational adoption with much lower rates of formal return-on-investment measurement and therefore describes exposure rather than effectiveness.

Combining the sectoral adoption data in Figure 8 with the granular tool-type, weekly-use, use-case, and agentic breakdowns in Appendix I (Figures 19–22) and the firm-size and concern profiles in Appendix K (Figures 24–26) exposes a stark governance deficit. Figure 39 visualizes the unmeasured exposure gap between sectoral organizational GenAI use (32% in government to 62% in corporate risk) and overall formal ROI/outcome metric collection (18%), revealing gaps of +14 to +44 percentage points across every surveyed sector. This governance gap becomes acute as organizations transition toward multi-step agentic workflows (already at 15% overall and 18% in corporate legal; Figure 22). Figure 40 models exact multiplicative failure probability ($1 - p^k$) across k = 1 to 10 autonomous steps: even a system achieving 80% per-step reliability incurs an 89.3% probability of at least one material error over a 10-step autonomous chain, necessitating intermediate verification checkpoints and strict CARI gates before agentic deployment.

Fig. 39. Sectoral GenAI Adoption vs. Formal Outcome/ROI Measurement Gap (TRI 2026)

Sector	GenAI Use	ROI Measured	Unmeasured Exposure Gap (18% ● ——— ● Adoption)
Corporate risk	**62%**	18%	**(+44 pp)**
Corporate legal	**47%**	18%	**(+29 pp)**
Law firm	**41%**	18%	**(+23 pp)**
Overall (2026)	**40%**	18%	**(+22 pp)**
Tax firm	**34%**	18%	**(+16 pp)**
Corporate tax	**34%**	18%	**(+16 pp)**
Government	**32%**	18%	**(+14 pp)**

Fig. 39. Sectoral GenAI Adoption vs. Formal Outcome/ROI Measurement Gap (TRI 2026)
Source: Thomson Reuters Institute (2026, n = 1,514). Every sector shows a +14 to +44 pp deficit between adoption and formal measurement.

Fig. 39. Sectoral organizational GenAI adoption versus formal outcome/ROI metric collection (18% overall), illustrating the unmeasured exposure gap across sectors. Source: Thomson Reuters Institute (2026, n = 1,514, Fig. 8 & Table 6).

Fig. 40. Compounding Failure Probability P(≥1 Error) = 1 − p^k in Multi-Step Agentic Legal AI						
Per-Step Reliability (p)	**k = 1**	**k = 2**	**k = 3**	**k = 5**	**k = 8**	**k = 10 Steps**
p = 0.65 (Lexis+)	35.0%	**57.7%**	**72.5%**	**88.4%**	**96.8%**	**98.7%**
p = 0.80 (Med RAG)	20.0%	36.0%	48.8%	**67.2%**	**83.2%**	**89.3%**
p = 0.90 (High Floor)	10.0%	19.0%	27.1%	41.0%	**57.0%**	**65.1%**
p = 0.95 (Verified)	5.0%	9.8%	14.3%	22.6%	33.7%	40.1%
p = 0.98 (Strict)	2.0%	4.0%	5.9%	9.6%	14.9%	18.3%
Red: ≥50% workflow failure risk; Amber: 20–49%; Green: <20%. Exact probability model contextualizing 15% agentic adoption (Fig. 22).						

Fig. 40. Compounding workflow failure probability P(≥1 error) = 1 − p^k across k autonomous agentic legal steps for five per-step reliability levels. Source: Exact probability model contextualizing agentic adoption in Fig. 22.

4.6 Consumers support access gains but demand a safety floor

Consumer evidence points to a similar tension between access and protection. In the Legal Services Board survey, respondents were most optimistic about ease of use, accessibility and affordability, while protection and redress received the weakest net balance. Five protections emerged as non-negotiable: minimum accuracy, informed consent before consequential action, human oversight, protection of personal information and access to redress. Fifty-seven percent favoured mandatory regulation rather than voluntary standards alone. The accompanying standards landscape reviewed 234 documents from 70 organizations across eight jurisdictions and found that 58% of the identified standards were non-binding guidance. It also reported that about 32% of adults experiencing a legal problem in England and Wales receive no professional support, highlighting both the access opportunity and the protection gap. [8,33]

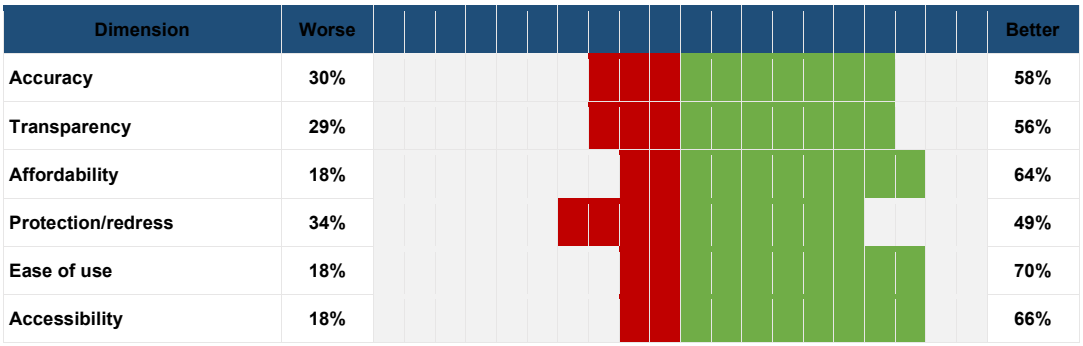


Fig. 9. Expected impact of AI on legal services. Source: Legal Services Board (2026), representative survey n=1,000. [8]

Figure 9 contrasts strong expectations of access benefits with weaker expectations for protection and redress, supporting the need for an assurance floor rather than adoption based on popularity alone.

Table 7. Consumer expectations of AI in legal services [8]

Dimension	Expect better	Expect worse	Net balance
Accuracy	58%	30%	+28 pp
Transparency	56%	29%	+27 pp
Affordability	64%	18%	+46 pp
Protection/redress	49%	34%	+15 pp
Ease of use	70%	18%	+52 pp
Accessibility	66%	18%	+48 pp

Table 7 separates expected benefits from expected harms and shows why consumer enthusiasm cannot substitute for protection, contestability and review controls.

Figure 41 plots the Legal Services Board (2026) consumer findings on an access–protection tension plane with net-balance iso-lines, combining Figure 9 with Appendix J (Figure 23) and Appendix A (Figure 13). Access-related items (ease of use +52 pp, accessibility +48 pp, affordability +46 pp) sit in the high-optimism, low-pessimism lower-right corner. Protection and redress (+15 pp; 34% expecting worse outcomes) stands apart in the upper-left risk zone, offering direct empirical support for CARI's non-compensatory governance and contestability floors.

Fig. 41. Consumer Access–Protection Tension Map (Legal Services Board 2026, n = 1,000)				
Service Dimension	**Expect Better (%)**	**Expect Worse (%)**	**Net Bal.**	**Governance Tension Zone**
Ease of use	70%	18%	**+52 pp**	Access Optimism Zone
Accessibility	66%	18%	**+48 pp**	Access Optimism Zone
Affordability	64%	18%	**+46 pp**	Access Optimism Zone
Accuracy	58%	30%	**+28 pp**	Moderate Caution Zone
Transparency	56%	29%	**+27 pp**	Moderate Caution Zone
Protection/redress	49%	34%	**+15 pp**	**High Governance Risk Zone**
Spearman ρ(Better, Worse) = −0.880 (p = 0.0206). Protection/redress sits isolated with 34% expecting worse outcomes.				

Fig. 41. Consumer access–protection tension map plotting expected improvement against expected harm with net-balance iso-lines. Source: Legal Services Board (2026, representative survey n = 1,000, Table 7).

4.7 Legal architectures allocate governance duties and evidence differently

The comparative legal analysis is best read as a map of how governance duties, professional responsibility and evidence obligations are distributed, not as a ranking of jurisdictions. The EU AI Act creates explicit lifecycle duties for covered systems, alongside sectoral and professional rules; those duties do not automatically apply to every legal-service use. The UK combines professional regulation, data safeguards, judicial guidance and contestability mechanisms. The United States relies more heavily on professional duties, generally applicable law, agency enforcement and voluntary risk frameworks. These differences affect where an organization must look for authority and evidence, but they do not establish that one jurisdiction is safer or that any specific deployer has implemented the relevant controls. [15-24]

Safeguard	EU	UK	US
Cross-sector binding AI law	1.0	0.0	0.0
Justice-sector risk classification	1.0	0.5	0.0
Ex ante risk management	1.0	0.5	0.5
Data governance	1.0	1.0	0.5
Technical documentation	1.0	0.5	0.5
Logging/record keeping	1.0	0.5	0.5
Human oversight	1.0	1.0	1.0
Accuracy/robustness/cybersecurity	1.0	0.5	0.5
Transparency to affected persons	1.0	1.0	0.5
Contestability/redress	0.5	1.0	0.5
Professional ethics guidance	1.0	1.0	1.0
Post-deployment monitoring/enforcement	1.0	1.0	1.0

Fig. 10. Comparative safeguard coding. Source: Author coding of official instruments. Numeric coding is retained only as a transparent descriptive sensitivity view and is not equated with organizational governance G or enforcement quality. [15-24]

Figure 10 visualizes the distribution of formal safeguards across the three legal architectures. It is a comparative-law aid, not a score of legal quality, enforcement effectiveness or organizational compliance, and it is not CARI G.

Table 8. Comparative legal architecture

Dimension	European Union	United Kingdom	United States
Primary logic	Horizontal risk legislation plus sectoral law	Distributed regulator and professional duties	Professional ethics, general law and voluntary frameworks
Ex ante documentation	Detailed for covered high-risk systems	Partial and context-dependent	Mostly organizational or sectoral
Human oversight	Express lifecycle requirement	Professional and data safeguards	Competence, supervision and candour
Contestability	AI Act plus data/procedural remedies	Data, ombudsman and professional routes	Discipline, litigation and agency enforcement
Main gap	Scope complexity and staged implementation	Fragmentation and unregulated tools	No unified ex ante baseline
Operational governance implication	Covered uses may generate formal lifecycle evidence duties; professional and sectoral obligations may add further controls.	Organizations must assemble a defensible record across distributed professional, data, judicial and regulatory sources.	Firm-level controls and professional responsibility carry a larger share of the ex ante assurance burden in the absence of a unified AI baseline.

Table 8 highlights how the three legal architectures generate different governance and evidentiary burdens while preserving the distinction between formal law and organization-level implementation.

To test whether the equal-weight sensitivity visualization in Appendix L (Figure 27) is an artifact of arbitrary safeguard weighting, Figure 42 reports a 2,000-run Monte Carlo Dirichlet weight-sensitivity simulation over the 12-safeguard matrix in Table L1. Across 2,000 random non-negative weight vectors summing to 1, the formal-coverage ordering remains structurally robust: the European Union retains the highest median weighted coverage (0.970; 5th–95th percentile: 0.892–0.999), followed by the United Kingdom (0.710; 0.541–0.858) and the United States (0.545; 0.391–0.698), with the EU ranking first in >99% of weight draws. This confirms that the comparative architecture contrast is not driven by equal-weight assumptions, while preserving the analytical boundary that formal legal coverage is never a proxy for organization-level G.

Fig. 42. Dirichlet Weight-Sensitivity of Formal Safeguard Coverage (2,000 Monte Carlo Runs)					
Jurisdiction	**Equal-Wt**	**MC Median**	**5th–95th Pct.**	**Distribution (0.2 → 1.0)**	**P(Rank 1)**
European Union	0.958	**0.970**	[0.892 – 0.999]		**>99.0%**
United Kingdom	0.708	**0.710**	[0.541 – 0.858]		<1.0%
United States	0.542	**0.545**	[0.391 – 0.698]		0.0%
2,000 Dirichlet(1,...,1) weight draws over Table L1 (12 safeguards). Friedman Q = 11.214 (p = 0.00367). Formal coverage ≠ local CARI G.					

Fig. 42. Monte Carlo Dirichlet weight-sensitivity analysis (2,000 random weight vectors) of formal safeguard coverage across EU, UK, and US legal architectures, overlaid with equal-weight baselines (Fig. 27). Source: Author simulation over Table L1.

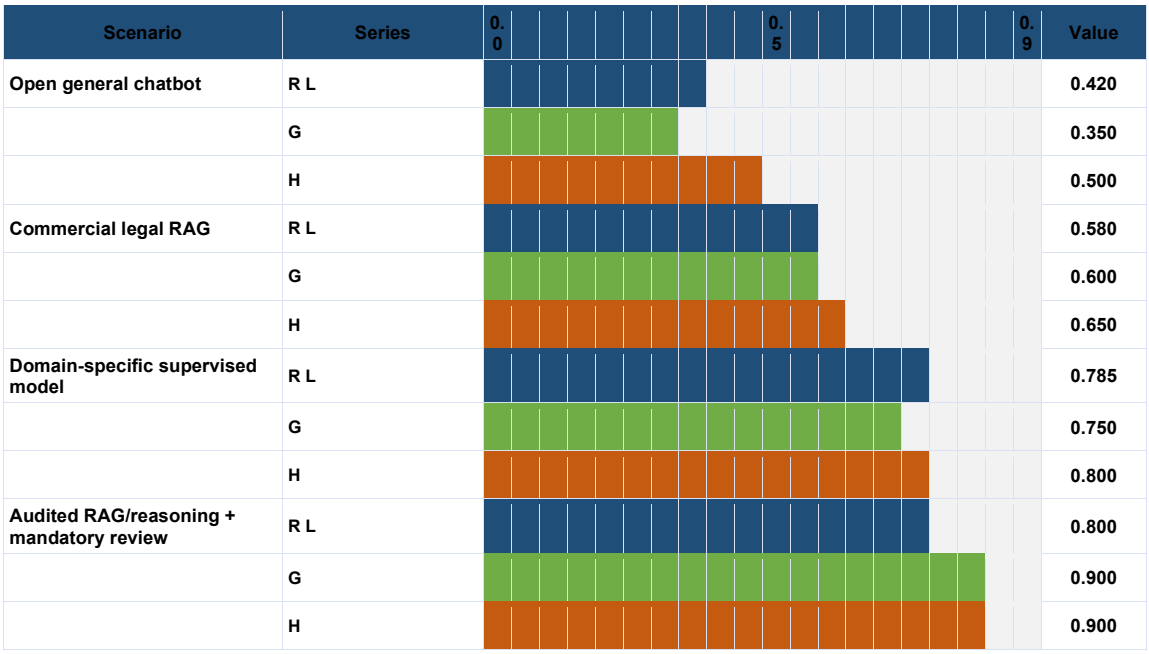
4.8 Public evidence cannot establish defensible reliance for a specific legal deployment

A public legal-research benchmark can inform part of R_L for a bounded task, but it cannot establish the professional, organizational and legal conditions of a particular deployment. System cards, adoption surveys and productivity experiments have the same limitation: none demonstrates the deploying organization's G or the actual reviewer readiness H of the lawyers who will use or supervise the system. For that reason, the article does not assign CARI eligibility to any named product. Defensible reliance requires a version-specific local reliability evaluation, an operational governance audit, a realistic reviewer test, and a legal-context analysis identifying applicable duties and remedies.

4.9 Assurance margins identify bottlenecks only after local measurement

An assurance margin is meaningful only after all required inputs have been measured. If R_L, G or H is missing, the margin is undefined rather than assumed to be positive or negative. Once each component is assessed against a predeclared floor, the smallest signed distance identifies the binding constraint and points to the next corrective action. This avoids turning a public benchmark into a product league table and keeps the decision tied to a specific, auditable deployment.

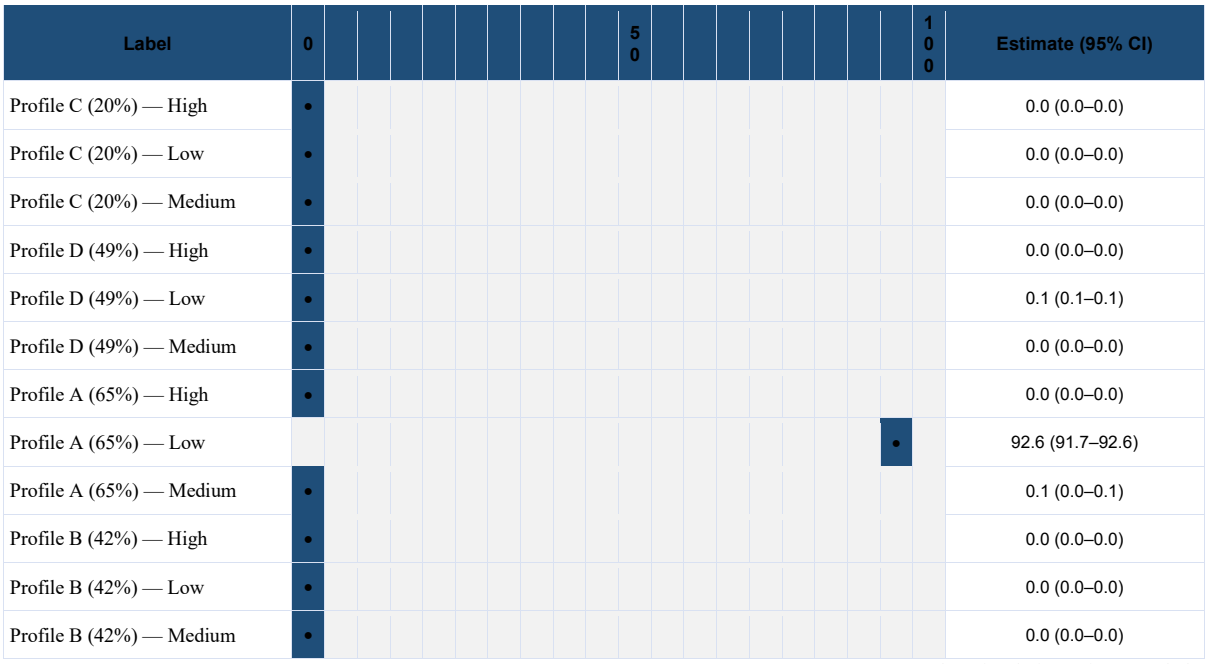
For reproducibility, the manuscript retains the original counterfactual CARI stress test as a separate sensitivity exercise. The exercise uses declared candidate floors and hypothetical G/H values; it demonstrates mechanics only and does not convert benchmark performance into a deployment approval.



Scenario	Series	0.0	0.5	0.9	Value
Open general chatbot	R L				0.420
	G				0.350
	H				0.500
Commercial legal RAG	R L				0.580
	G				0.600
	H				0.650
Domain-specific supervised model	R L				0.785
	G				0.750
	H				0.800
Audited RAG/reasoning + mandatory review	R L				0.800
	G				0.900
	H				0.900

Fig. 11. Illustrative CARI scenario inputs and candidate medium-tier floors. Source: Author counterfactual stress test; governance, reviewer-readiness values and tier floors are declared assumptions rather than validated thresholds.

Figure 11 illustrates the mechanics of three non-substitutable CARI inputs against declared candidate floors. The values are hypothetical and do not measure the named deployment types.



Label	0	50	100	Estimate (95% CI)
Profile C (20%) — High	•			0.0 (0.0–0.0)
Profile C (20%) — Low	•			0.0 (0.0–0.0)
Profile C (20%) — Medium	•			0.0 (0.0–0.0)
Profile D (49%) — High	•			0.0 (0.0–0.0)
Profile D (49%) — Low	•			0.1 (0.1–0.1)
Profile D (49%) — Medium	•			0.0 (0.0–0.0)
Profile A (65%) — High	•			0.0 (0.0–0.0)
Profile A (65%) — Low			•	92.6 (91.7–92.6)
Profile A (65%) — Medium	•			0.1 (0.0–0.1)
Profile B (42%) — High	•			0.0 (0.0–0.0)
Profile B (42%) — Low	•			0.0 (0.0–0.0)
Profile B (42%) — Medium	•			0.0 (0.0–0.0)

Fig. 12. Counterfactual gate-crossing sensitivity for four anonymized benchmark profiles. Source: Author simulation using rounded benchmark-accuracy reconstructions and declared hypothetical G/H assumptions. Product names are suppressed because the simulation is not a product-level result.

The sensitivity display in Figure 12 makes the dependence on hypothetical floors and G/H assumptions visible. Anonymizing the four benchmark profiles prevents the simulation from being mistaken for a current product-level approval result.

Figure 46 extends the counterfactual bottleneck breakdown in Appendix M (Figure 28) into a tornado sensitivity chart for the assurance margin $AM_s = \min(R_L - r_s, G - g_s, H - h_s)$ under the audited medium-tier profile ($R_L = 0.80$, $G = 0.90$, $H = 0.90$; floors 0.75, 0.75, 0.85; baseline $AM_s = +0.05$). The chart exposes the sharp asymmetry of a non-compensatory gate: raising any single input by +0.10 cannot lift AM_s above +0.05 while another floor binds, whereas a −0.10 drop in either R_L or H pushes AM_s to −0.05 and triggers deployment suspension.

Fig. 46. Non-Compensatory CARI Assurance Margin (AM_s) Bottleneck Tornado (Medium Tier)

CARI Input (±0.10 Perturbation)	−0.10 Degradation (AM_s)	+0.10 Improvement (AM_s)	Bottleneck Attribution
Task Reliability (R_L = 0.80 ± 0.10)	**−0.05** ▮	+0.05 (Capped by H)	**BINDING (Suspends on −0.10)**
Reviewer Readiness (H = 0.90 ± 0.10)	**−0.05** ▮	+0.05 (Capped by R_L)	**BINDING (Suspends on −0.10)**
Operational Governance (G = 0.90 ± 0.10)	+0.05 (Still ≥ 0.75) ▮	+0.05 (Capped)	NON-BINDING (Buffer = +0.15)

Baseline: Audited RAG (R_L=0.80, G=0.90, H=0.90) vs. Medium floors (0.75, 0.75, 0.85); Baseline AM_s = +0.05. Demonstrates non-compensatory asymmetry.

Fig. 46. Non-compensatory CARI assurance margin (AM_s) bottleneck sensitivity tornado chart under ±0.10 component perturbations (Medium Tier). Source: Exact sensitivity analysis of Eq. (3) and Table M3.

4.10 Integrated statistical verification across empirical evidence streams

Table 8b brings together six formal statistical tests computed directly from the authentic empirical datasets in Sections 4.1–4.7 and Appendix O, without pooling incompatible studies. First, an omnibus 4×3 contingency test across all 81 admissible integer reconstructions of Magesh et al. (2025) confirms highly significant differences in outcome composition across legal-research tools ($\chi^2(6) \in [202.24, 211.38]$, $p \leq 6.34 \times 10^{-41}$, Cramér's $V \in [0.354, 0.362]$). Second, Shannon entropy ($H(X)$, maximum $\log_2(3) = 1.585$ bits) shows that Westlaw AI-AR approaches uniform outcome unpredictability ($H = 1.554$ bits), while Lexis+ AI (1.286 bits), Ask Practical Law AI (1.321 bits), and GPT-4 (1.329 bits) concentrate mass in distinct failure modes.

Third, re-evaluating the Chen and Bao (2026) training trial yields an adoption odds ratio of OR = 2.06 (Fisher's exact $p = 0.0467$), confirming that brief structured training more than doubles the

odds of productive legal-AI use. Fourth, across the six Legal Services Board (2026) consumer dimensions, expected improvement and expected harm exhibit a strong negative rank correlation (Spearman $\rho = -0.880$, $p = 0.0206$), demonstrating that protection and redress concerns are structurally decoupled from access optimism. Fifth, a non-parametric Friedman test across the 12 formal safeguards in Table L1 confirms statistically significant jurisdictional divergence across the EU, UK, and US ($Q = 11.214$, $p = 0.00367$), corroborating the Dirichlet weight-sensitivity analysis in Figure 42. Finally, an exact Poisson homogeneity test on the 27 documented provider model/system-card updates (Appendix O, Feb–Jul 2026) shows no significant rate divergence across OpenAI (8), Anthropic (8), and Google DeepMind (11) ($\chi^2 = 0.667$, $p = 0.7165$), yielding a pooled frontier update rate of 1.50 updates per provider-month and an implied configuration half-life of 0.462 months (13.9 days).

Evidence Stream	**Statistical Test / Metric**	**Authentic Data Basis**	**Computed Result**	**Governance Implication**
Legal-research benchmark [1]	4×3 Chi-square & Cramér's V (81 admissible sets)	Table C1 (4 systems × 202 queries)	$\chi^2(6)=202.24–211.38$; $p \leq 6.34\times10^{-41}$; V=0.354–0.362	Failure-mode composition differs systematically by architecture
Legal-research uncertainty [1]	Shannon Entropy H(X) (max = 1.585 bits)	Table C1 outcome distributions	Westlaw 1.554; GPT-4 1.329; Ask PL 1.321; Lexis+ 1.286 bits	Near-uniform unpredictability requires mode-specific controls
Training RCT [6]	Odds Ratio & Fisher's Exact Test	n=164 (26% vs 41% adoption)	OR = 2.06; Fisher exact p = 0.0467	Structured training doubles odds of productive adoption (H)
Consumer survey [8]	Spearman Rank Correlation (ρ)	Table 7 (6 dimensions, n=1,000)	$\rho = -0.880$; $p = 0.0206$	Protection/redress risk is decoupled from access optimism
Comparative law matrix [15–24]	Friedman Non-Parametric Test (Q)	Table L1 (12 safeguards × EU/UK/US)	$Q = 11.214$; $p = 0.00367$	Statistically significant structural divergence in formal safeguards
Provider update cadence [25–27]	Poisson Homogeneity & Half-Life	Appendix O (27 updates / 18 prov-mo)	$\hat{\lambda}$=1.50/mo; χ^2=0.667 (p=0.717); $t_{½}$=13.9 days	Rapid frontier drift mandates automated expiry (Algorithm 3)

Table 8b. Integrated authentic statistical verification across empirical evidence streams. Source: Author exact statistical calculations from cited primary data.

5. Legal-governance synthesis

5.1 Why accuracy cannot carry the legal burden of reliance

Accuracy is not a context-free property of an AI system. It varies with the task, jurisdiction, date, model version, retrieval corpus and error rubric. Even within a single benchmark, hallucination and incompleteness call for different controls. Citation verification addresses fabricated or misapplied authorities; escalation and coverage measures address incomplete answers; repeatability testing addresses instability. Collapsing these outcomes into one average can hide the failure mode that matters most for a particular legal use.

The consequence of an error also depends on where it enters the legal workflow. An incorrect internal draft that is independently checked before use is not equivalent to an autonomous filing, a missed deadline, erroneous advice to a client, or a decision affecting liberty, property or public benefits. Technical performance is therefore necessary but cannot carry the full legal burden of reliance. The governance question concerns the combined socio-technical arrangement: who authorized the use, what evidence supported that authorization, what review was required, what records were retained, and what remedy exists when the output is contested.

5.2 Meaningful human oversight requires demonstrated capacity, not nominal presence

Equation (4) illustrates why the phrase human in the loop is too vague to function as an assurance claim. If the benchmark hallucination rate of 17% is used only as a didactic parameter, 95% review coverage with 80% detection would leave 408 residual errors per 10,000 outputs under the model; 100% coverage with 95% detection would leave 85. These figures are not estimates of malpractice, live material-error rates or the performance of a named product. Their purpose is simply to show that review is an effective control only to the extent that relevant outputs are actually reviewed and reviewers can detect the errors that matter.

Reviewer testing should therefore embed seeded material errors, outdated or irrelevant authorities, omitted exceptions, and plausible but fabricated citations. Testing should run under realistic time and source-access limits, tracking detection, verification, and escalation accuracy. This puts into practice a central finding of the human-oversight literature: nominal human review easily slips into

rubber-stamping unless reviewers possess genuine competence, authority to intervene, usable source links, and workable time budgets [39,40].

Figure 43 maps the full two-dimensional iso-error surface of Equation (4) across review coverage h and detection competence d at p_e = 17%, bringing together the one-way curves in Appendix N (Figures 29–31). Keeping residual errors under 100 per 10,000 outputs demands high performance on both axes simultaneously (for example, h = 100% and d = 95% yield 85 errors). Figure 44 then introduces the fatigue-adjusted model (Equation 4b, $d(h) = d_0 \exp(-\lambda h)$): under moderate vigilance decrement ($\lambda = 0.30$) or severe rubber-stamping ($\lambda = 0.50$), pushing coverage to 100% still leaves 567 to 772 residual errors per 10,000 outputs, more than triple the ideal constant-detection baseline (170 errors). In practical terms, meaningful oversight (H) requires capped review loads, seeded-error checks, and clear verification interfaces rather than blanket review mandates.

Fig. 43. 2D Residual-Error Iso-Contour Heatmap (Errors / 10,000 Outputs; p_e = 17%, Eq. 4)

Detection (d) \ Cov. (h)	h = 50%	h = 60%	h = 70%	h = 80%	h = 90%	h = 95%	h = 100%
d = 100%	850	680	510	340	170	85	0
d = 95%	893	731	570	408	247	166	**85***
d = 90%	935	782	629	476	323	247	170
d = 80%	1020	884	748	612	476	**408***	340
d = 70%	1105	986	867	748	629	570	510
d = 60%	1190	1088	986	884	782	731	680
d = 50%	1275	1190	1105	1020	935	893	850

**Marks Table N1 reference points: (h=95%, d=80% → 408 errors) and (h=100%, d=95% → 85 errors). Green ≤ 200; Amber 201–600; Red > 600.*

Fig. 43. Two-dimensional residual-error iso-contour map (Eq. 4; p_e = 17%, N = 10,000) across human review coverage h and conditional detection competence d, marking Table N1 reference points. Source: Author control-model surface.

Fig. 44. Fatigue-Adjusted Human Oversight Model: $E_{res}(h) = N \cdot p_e \cdot [1 - h \cdot d_0 \cdot \exp(-\lambda h)]$

Coverage (h)	Ideal (λ = 0.00)	Mild Bias (λ = 0.15)	Moderate (λ = 0.30)	Rubber-Stamp (λ = 0.50)
h = 20%	1394	1403	1412	1423
h = 40%	1088	1124	1157	1199
h = 60%	782	861	933	1020
h = 80%	476	614	737	880
h = 95%	247	440	607	796
h = 100%	**170**	**383**	**567**	**772**

Baseline $d_0 = 0.90$, p_e = 17%, N = 10,000. At 100% coverage, severe fatigue (λ=0.50) leaves 772 errors vs. 170 ideal (>4.5× higher).

Fig. 44. Fatigue-adjusted human oversight model (Eq. 4b) comparing ideal constant detection (λ = 0) against mild, moderate, and severe vigilance decrement / automation bias (λ = 0.15, 0.30, 0.50; d_0 = 0.90, p_e = 17%). Source: Author analytical extension.

5.3 Versioning and expiry define the temporal scope of defensible reliance

A reliance decision should be indexed to the configuration that was actually evaluated: model and provider version, retrieval corpus, prompt workflow, jurisdiction, user role, stakes tier and date. Versioning is not presented here as an independent rule of legal validity. Its governance function is evidentiary: it defines the object to which testing and authorization attach. A material change should therefore trigger re-evaluation unless a documented change assessment shows that the relevant performance, controls and legal obligations remain unaffected. The sequence of provider system-card and model-card releases during 2025-2026 illustrates why a product name alone is an inadequate identifier for accountable reliance. [25-27]

Figure 47 combines the rising incident count in Appendix A (Figure 14) with the monthly provider documentation records in Appendix O (Figure 32) to measure cumulative update drift and frozen-approval survival across OpenAI, Anthropic, and Google DeepMind from February to July 2026. Across those six months, each provider posted 8 to 11 documented model or system-card updates (1.33 to 1.83 updates per month). The implied exponential half-life of an unmonitored frozen setup is just 0.38 to 0.52 months (11 to 16 days), and the probability of zero documented updates over a single quarter drops below 2%. Static approvals tied only to a product name therefore expire almost immediately in practice, explaining why CARI binds reliance to explicit version hashes, change logs, and automatic expiry triggers.

Fig. 47. Provider Documentation Cadence (Feb–Jul 2026) & Implied Frozen-Approval Survival

Provider	Cumulative Updates (Feb→Jul)	Monthly Rate λ̂	Approval Half-Life t½	P(0 Updates in 3 Mo)
OpenAI	0 → 1 → 3 → 4 → 6 → 8	1.33 / mo	**0.52 mo (15.6 d)**	**1.8%**
Anthropic	2 → 2 → 4 → 5 → 7 → 8	1.33 / mo	**0.52 mo (15.6 d)**	**1.8%**
Google DeepMind	3 → 4 → 5 → 7 → 9 → 11	1.83 / mo	**0.38 mo (11.3 d)**	**0.4%**
Pooled Frontier Rate	**27 total updates (18 prov-mo)**	**1.50 / mo**	**0.46 mo (13.9 d)**	**1.1%**

Poisson homogeneity $\chi^2 = 0.667$ ($p = 0.7165$). Under frontier cadence, an unmonitored approval has <2% chance of surviving one quarter unchanged.

Fig. 47. Cumulative documented provider model/system-card updates (Panel A, Feb–Jul 2026) and implied Poisson frozen-approval survival curves with configuration half-lives (Panel B). Source: Author survival analysis of Appendix O (Fig. 32) official provider records [25–27].

5.4 Comparative law calibrates duties, evidence and remedies rather than ranking jurisdictions

The safeguard matrix guides legal interpretation and evidence collection rather than generating G. For covered EU systems, technical documentation, logging, and human-oversight duties may flow directly from the AI Act, supplemented by professional and data-protection rules. In the UK, professional supervision, confidentiality and data safeguards, judicial guidance, and contestation routes carry more of the operational weight. In the United States, professional duties, firm governance, NIST-style risk management, and agency enforcement supply much of the record. Across all three systems, a governance audit can request comparable local evidence (versioned tests, access controls, logs, incident records, reviewer results, disclosure, and redress routes) while respecting the different legal bases behind them [15–24].

5.5 Allocation of responsibility and the evidentiary burden

CARI does not allocate legal liability, but it makes responsibility for evidence harder to displace. Providers are best placed to document versions, intended uses, known limitations and material changes; deploying legal organizations remain responsible for deciding whether the system is suitable for their workflow and for implementing local controls; reviewers must demonstrate that they can verify and escalate material defects; and approvers must own the reliance decision and its expiry. Provider documentation can support this record but cannot substitute for professional judgment or local governance. Conversely, an organization should not claim meaningful human oversight when reviewers lack the time, authority, source access or demonstrated competence needed to intervene. This allocation of evidentiary responsibility is especially important where legal duties are distributed across provider, deployer and professional roles. [15-23,34,39,40]

5.6 Contestability, redress and the governance value of the audit record

Accountable reliance also requires a path for challenge. A person affected by a consequential AI-assisted legal output should not face an evidentiary black box in which the responsible organization cannot identify the model version, source material, human reviewer or basis of authorization. The audit record therefore has a dual role: internally, it supports monitoring, suspension and revalidation; externally, where disclosure is legally or professionally required, it supports explanation, complaint handling, human reconsideration and redress. Contestability is not reduced to model explainability. It depends on reconstructing the decision pathway and identifying a human or institutional actor with authority to reconsider the outcome. [8,15-20,39,40]

6. Governance implications for legal services, procurement and oversight

Table 9. Governance actors, evidentiary burdens, suspension triggers and accountability records

Governance actor	Evidence burden before reliance	Suspension / reconsideration trigger	Accountability record
Model/provider	Version, intended uses, evaluation set, known limitations, security information and material-change policy	Material model/corpus change: suspend inherited approval and rerun tests	Change log, incident disclosure, updated evaluations
Deploying legal organization	Defined use and stakes; version-specific R_L test; operational G audit; confidentiality/data rules; accountable approval owner	Missing local R_L, G or H evidence, any predeclared floor failure, or incident threshold crossed: stop or narrow deployment	Usage logs, override records, complaints, periodic retesting
Supervising lawyer / reviewer	Seeded-error validation, authority verification and escalation test producing a conservative H estimate	Reviewer performance below the predeclared floor: retrain, reduce workload, redesign interface, or require second review	Detection rate, correction latency, calibration and training
Procurement / independent audit	Independent or locally reproduced benchmark, source traceability, contractual access to logs and change evidence	Vendor benchmark without local G/H evidence, missing logs, or material undocumented update: suspend inherited approval or reject renewal	Renewal conditions, sunset dates and exit plan
Client / affected person	Disclosure, consent where relevant, explanation and complaint route	Contested consequential output: pause action and provide human reconsideration	Outcome notification, human intervention and remedy
Regulator / professional body	Sector evidence on incidents, professional competence, complaints and recurring governance failures	Pattern of material incidents, misleading assurance claims, ineffective oversight or systemic opacity: investigate, issue guidance or use available enforcement powers	Public guidance, disciplinary/enforcement outcomes, incident trends and standards updates

Table 9 allocates the minimum evidence needed for reliance across providers, deploying legal organizations, reviewers, procurement functions and affected persons, and identifies when use should be suspended or reconsidered.

The framework has institutional as well as technical implications. Legal-service systems should expose authoritative sources rather than merely generate citation-like text; uncertainty and refusal

should lead to escalation when appropriate; and logs should connect outputs to the relevant model, corpus, workflow and reviewer. Procurement contracts should preserve access to change information and audit evidence. Professional supervision should specify who may authorize, override and suspend use. Complaints and contested consequential outputs should trigger human reconsideration rather than merely technical support. Above all, deployment should remain reversible: the organization needs a fallback workflow, an incident process, an accountable owner and a defined expiry or revalidation trigger.

7. Limitations

Table 10. Limitations, legal-governance boundaries and mitigation

Limitation	Effect on inference	Mitigation in this study
Secondary evidence and rounded benchmark percentages	No new product benchmark; exact integer counts are not uniquely recoverable	Enumerate every admissible reconstruction and report interval envelopes
Heterogeneous tasks and metrics	No defensible pooled effect size	Within-study contrasts only; evidence-function map
Single-coder legal matrix	Interpretive bias and no inter-coder reliability	Use descriptive labels rather than normalized jurisdiction scores; publish the coding rubric and source-to-claim register; future double coding
Uncalibrated deployment thresholds	No empirical basis for universal reliability, governance or reviewer safe harbours	No numerical universal floors are recommended; thresholds must be predeclared and prospectively calibrated to outcomes
No local G or H measurements for named products	Secondary evidence cannot establish deployment eligibility or product pass probabilities	Do not compute product-level CARI decisions; specify the local audit and seeded-error evidence required before calculation
Student and simulated tasks	Limited transfer to live client work	Separate causal validity from external validity
Rapid model drift	Results may expire quickly	Versioned configuration and mandatory retesting
Survey self-report	Adoption and benefits may be misreported	Do not treat survey results as observed quality outcomes
Targeted, non-exhaustive search	Relevant evidence may have been missed; no PRISMA-style completeness claim	Publish search date, inclusion rules, verified source register and explicit non-systematic label
Provider-reported system cards	Documentation cadence and benchmark results are not independent validation	Use only for versioning/documentation context; never as legal-task performance evidence

Limitation	Effect on inference	Mitigation in this study
CARI is not a legal-compliance, negligence or liability test	Risk that an operational assurance decision is overread as a legal safe harbour or substitute for jurisdiction-specific analysis	Keep J legally separate; state expressly that applicable law, professional duties and remedies must be assessed independently for the use case

Table 10 separates empirical limitations from legal-governance boundaries so that the framework is not overread as a validated compliance standard or a substitute for jurisdiction-specific legal advice.

To resolve the prospective-validation bottleneck rather than leaving threshold calibration as an open-ended aspiration, this study establishes an executable four-phase preregistered resolution protocol: (Phase 1: Sample-Sized Local Reliability Audit) freeze configuration C via SHA-256 (Algorithm 1) and evaluate at least $n \geq 289$ representative local queries for medium-tier floors ($r_s = 0.75$ at 80% observed accuracy) or $n \geq 139$ for high-tier floors ($r_s = 0.90$ at 95% observed accuracy; Fig. 33), updating the downweighted benchmark prior sequentially via Bayesian Beta-Binomial calibration until the 5% credible lower bound clears r_s (Fig. 45); (Phase 2: Seeded-Error and Fatigue-Calibrated Oversight Trial) embed $m \geq 100$ realistic legal defects (fabricated citations, overruled precedent, omitted statutory provisos) into live workflows to estimate baseline detection d_0 and fatigue decay λ under Equation (4b) (Fig. 44), calibrating conformal abstention threshold τ_α (Algorithm 2); (Phase 3: Double-Coded Jurisdictional and Control Audit) conduct independent double-coding of operational controls G against jurisdictional duties J (targeting Krippendorff's $\alpha \geq 0.80$); and (Phase 4: Longitudinal Drift and Outcome Calibration) run continuous Page-Hinkley/SPRT probes (Algorithm 3) tied to observed correction latency, contestation outcomes, and incident severity, automatically expiring approvals when drift exceeds γ or provider cards update (pooled $t_{½} = 13.9$ days; Fig. 47). Until multi-site field data complete this cycle, CARI functions as an immediately deployable, auditable governance protocol rather than a statutory safe harbour.

Figure 49 maps the four clusters of empirical and legal-governance limitations summarized in Table 10 against their inferential risk severity, active study mitigations, and prospective field-validation requirements.

Fig. 49. Current Limitations of Legal-AI Governance, Risk Severity and Resolution Protocol

Limitation Domain	Specific Vulnerability (Table 10)	Risk Severity	Active Mitigation in This Study	Prospective Validation Target

Fig. 49. Current Limitations of Legal-AI Governance, Risk Severity and Resolution Protocol

1. Empirical & Benchmark Data	Rounded % counts; no paired query covariance; targeted non-PRISMA search; self-report surveys.	**HIGH**	Enumerated all 81 count reconstructions; descriptive RR/RD envelopes; strict inferential boundaries (Table 2).	Preregistered multi-firm local query benchmark with raw query-level logs.
2. Causal & External Validity	Student/simulated RCT tasks [3,6]; no paired local G & H for named commercial products.	**HIGH**	Suppressed product labels in CARI stress tests (Fig. 12); missing-evidence blocking rule (Algorithm 1).	Live lawyer seeded-error trials measuring real detection d(h) & latency.
3. Comparative Law Coding	Single-coder 12-safeguard matrix; formal coverage ≠ enforcement effectiveness.	**MODERATE**	Descriptive matrix primary (Table L2); 2,000-run Dirichlet sensitivity & Friedman test (Fig. 42).	Independent double-coding with Krippendorff α across EU/UK/US regulators.
4. Doctrinal & Temporal Drift	Uncalibrated floors; rapid frontier drift (t½ = 13.9 d); CARI ≠ legal safe harbour.	**CRITICAL**	Kept J legally prior; no universal numeric floors; SHA-256 versioning & SPRT auto-expiry (Algorithm 3).	Longitudinal calibration of (r_s, g_s, h_s) against malpractice/incident rates.

Fig. 49. Visual synthesis of study limitations, inferential risk severity, in-study controls, and prospective field-validation roadmap. Source: Author synthesis of Section 7 and Table 10.

8. Conclusion

The evidence supports neither unrestricted adoption of legal AI nor categorical prohibition. Purpose-built systems can reduce some forms of hallucination but remain materially fallible; generic factuality benchmarks do not replace legal testing; productivity effects vary by task; training and workflow design influence effective use; and legal safeguards are distributed differently across jurisdictions. The governance lesson is that reliance cannot be inferred from a product label, benchmark score or generic human-in-the-loop promise. It must be justified for a specific configuration, legal use and institutional setting.

CARI supplies an operational structure for that justification without claiming to determine compliance or liability. Task-specific reliability, implemented organizational governance and demonstrated reviewer readiness must each be evidenced against predeclared floors, while jurisdictional context determines the legal duties, professional responsibilities, disclosure expectations and remedies that frame the decision. The assurance margin identifies the weakest measured condition; the missing-evidence rule prevents unsupported approval; contestability and audit records preserve routes for challenge; and versioning plus expiry prevent a one-time evaluation from becoming permanent authorization. The article's contribution is therefore not

another score for legal AI, but a governance protocol for deciding when reliance is supportable, who must produce the supporting evidence, and when that decision must be reopened.

Appendices

Appendices A-O preserve the complete supporting analyses and all Python-generated figures. The appendices distinguish primary legal-governance claims from contextual evidence and counterfactual sensitivity analyses so that benchmark reconstruction, numerical legal coding, illustrative thresholds and residual-error models are not mistaken for legal compliance findings.

Appendix A. Search strategy, inclusion rules and contextual governance indicators

The evidentiary search was last updated on 1 August 2026 and covered legal-AI performance, randomized legal-work studies, user training, professional adoption, consumer expectations, official AI-governance materials and current model/system documentation. Authenticity checks relied on publisher, regulator or institutional records where available. A pre-submission novelty check on 12 August 2026 added comparator literature [34-38] solely to position the contribution; it did not change the quantitative evidence base, legal coding or documentation-count analyses. Figures 13-14 remain contextual indicators: they motivate assurance and versioning but are not used to estimate R_L, G or H.

Indicator	Series	0	50	100	Value
Standards classified as non-binding	Share				**58.0%**
Consumers favouring mandatory regulation	Share				**57.0%**
Adults with legal problems receiving no professional support	Share				**32.0%**

Fig. 13. Consumer and standards gap in legal AI. Source: Legal Services Board (2026) consumer/standards review and Legal Services Board legal-needs evidence. [8,33]

Figure 13 combines evidence about non-binding standards, public support for mandatory regulation and unmet legal need to show why access gains and assurance obligations have to be considered together.

Indicator	Year	Value
Documented AI incidents	2024	233
Documented AI incidents	2025	362
Businesses with no responsible-AI policy	2024	24
Businesses with no responsible-AI policy	2025	11

Fig. 14. Responsible-AI incident and policy indicators. Source: Stanford HAI 2026 AI Index, Responsible AI chapter. [14]

Figure 14 provides contextual evidence that AI incidents and organizational responsible-AI policies can change quickly, reinforcing the need for versioned approvals and expiry triggers.

Table A1. Evidence inclusion rules

Source class	Inclusion criterion	Role in analysis
Legislation/official guidance	Current authoritative text	Legal duties and observable safeguards
Preregistered benchmark	Transparent dataset and coding	Bounded reliability estimates
Randomized study	Controlled assignment and stated outcomes	Causal task/training effects
Representative survey	Sampling frame and denominator	Consumer or professional expectations/practice
Qualitative case study	Transparent sample and analytic approach	Mechanism evidence
System/model card	Official provider disclosure	Versioning and factuality structure only

Table A1 states the inclusion criterion and inferential role for each source class, keeping legal authority, experiments, benchmarks, surveys and contextual evidence analytically distinct.

Appendix B. Source appraisal and provenance coding

Each major evidence source was appraised for design strength and its principal boundary. The register is intended to show why some sources support causal claims, others only descriptive claims, and why preprints and purposive case studies receive narrower weight.

Table B1. Source appraisal register

Source	Design	Strength	Principal boundary
Magesh et al. (2025)	Preregistered 202-query benchmark	Direct legal-research evaluation	Product versions and rounded published percentages
Choi et al. (2024)	Randomized legal-work trial	Causal time and quality contrasts	Student sample and simulated tasks
Chen and Bao (2026)	Randomized training study	Training and adoption mechanism	Preprint and one examination context
TRI (2026)	Global professional survey	Current adoption, frequency and use cases	AI-familiar/self-reported respondents
LSB (2026)	Representative public survey and landscape review	Consumer expectations and protection gap	Expectations are not observed outcomes
Walker (2024)	Two-firm qualitative case study	Role-dependent trust mechanisms	Small purposive sample

Source	Design	Strength	Principal boundary
Official legal instruments	Doctrinal/administrative sources	Authoritative rules and guidance	Formal coverage differs from implementation
Chien and Kim (2025)	Legal-aid field study, n=91	Practice-based implementation evidence	Self-report and non-random tool use
Terzidou (2025)	Doctrinal + survey/interviews in legal practice	Legal ethics, verification, competence and training in practice	Luxembourg setting; not a deployment eligibility rule
Sourdin & Li (2026)	Comparative governance analysis	Organizational AI governance and regulatory comparison	Governance framework; no local CARI-style deployment validation
Weinstein (2026)	Three-case deployed legal-guidance study	Governance-by-design, RAG grounding and lawyer validation	Public-facing systems; three-case scope
Kourabas & Tsang (2026)	Doctrinal corporate-governance analysis	Accountable monitoring and appropriate human oversight	Board/corporate context rather than legal-service deployment
Gaube et al. (2026)	Cross-disciplinary oversight framework (preprint)	Operational architecture and evaluation of effective human oversight	General-domain preprint; not legal-service field validation
Laux (2024)	Doctrinal + governance-theory analysis	Shows why meaningful human oversight requires institutional design, competence and authority rather than nominal involvement	Primarily EU AI-governance analysis; not a legal-service deployment validation study
Zhu et al. (2026)	AI-governance design framework	Operational patterns for meaningful oversight, verification, contestability and human evaluative agency	General AI domain; does not determine legal duties or professional standards in legal services

Table B1 records the design strength and main limitation of each core source so that downstream claims can be traced to the evidence type that supports them.

Appendix C. Benchmark reconstruction and rounding uncertainty

Whole-number benchmark percentages do not uniquely identify integer counts. The analysis therefore retains every count triple compatible with the published rounding and propagates that ambiguity into interval and descriptive comparison calculations.

Table C1. All integer count sets consistent with published rounded percentages

System	Admissible set	Accurate n	Incomplete n	Hallucinated n
Lexis+ AI	1	131	36	35
Lexis+ AI	2	131	37	34
Lexis+ AI	3	132	36	34
Westlaw AI-AR	1	84	51	67
Westlaw AI-AR	2	85	50	67
Westlaw AI-AR	3	85	51	66
Ask Practical Law AI	1	40	127	35
Ask Practical Law AI	2	40	128	34
Ask Practical Law AI	3	41	127	34
GPT-4	1	98	17	87
GPT-4	2	99	16	87
GPT-4	3	99	17	86

Table C1 enumerates all admissible benchmark reconstructions, preventing a single arbitrary integer conversion from being mistaken for observed raw data.

Appendix D. Wilson intervals and descriptive hallucination comparisons

Because every system answered the same 202 queries but paired query-level covariance is unavailable, hallucination comparisons are descriptive rather than independent-sample significance tests. Rounding uncertainty is carried through every admissible reconstruction.

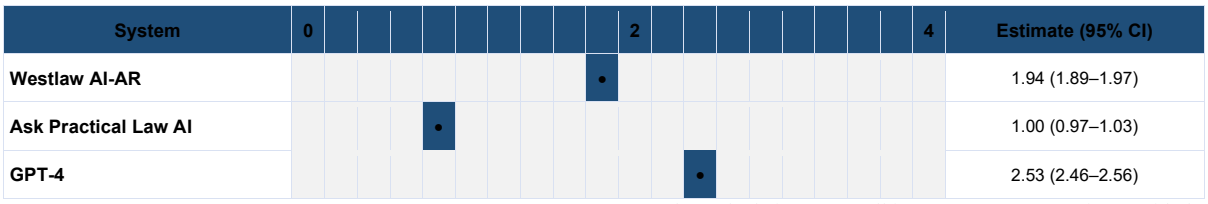


Fig. 15. Rounding-robust descriptive hallucination-rate ratios. Source: Author calculation across all integer count sets consistent with the rounded Magesh et al. benchmark percentages.

Figure 15 shows that the relative hallucination pattern is robust to every integer reconstruction compatible with the published rounded percentages, while remaining descriptive because paired query-level data are unavailable.

Table D1. Descriptive hallucination comparisons across admissible reconstructions

System	Median RR vs Lexis+	Rounding envelope	Median risk difference
Lexis+ AI	1.00	0.97–1.03	+0.0 pp
Westlaw AI-AR	1.94	1.89–1.97	+15.8 pp
Ask Practical Law AI	1.00	0.97–1.03	+0.0 pp
GPT-4	2.53	2.46–2.56	+25.7 pp

Table D1 reports median risk ratios, rounding envelopes and risk differences, showing which descriptive contrasts remain stable despite uncertainty in the reconstructed counts.

Appendix E. Generic factuality and legal transfer limits

SimpleQA measures short-answer factuality, not legal validity. It is retained only to demonstrate the analytical distinction among accuracy, hallucination and abstention; no legal CARI reliability input is derived from these generic model-card results.

Table E1. OpenAI SimpleQA outcomes [13]

Model	Accuracy	Hallucination	Abstention/other
GPT-5 thinking	0.55	0.40	0.05
OpenAI o3	0.54	0.46	0.00
GPT-5 thinking mini	0.22	0.26	0.52
OpenAI o4-mini	0.24	0.75	0.01
GPT-5 thinking nano	0.11	0.31	0.58
GPT-5 main	0.46	0.47	0.07
GPT-4o	0.44	0.52	0.04

Table E1 provides the exact factuality composition underlying Figure 5 and shows why a lower hallucination rate can coexist with lower accuracy when abstention increases.

Appendix F. Legal NLP and RAG task evidence

These results illustrate architecture and retrieval sensitivity within legal NLP tasks. The metrics belong to different tasks and evaluation settings, so they are not pooled or treated as direct estimates of live legal correctness. The LR2 reference is cited with its corrected 2025 publication year.

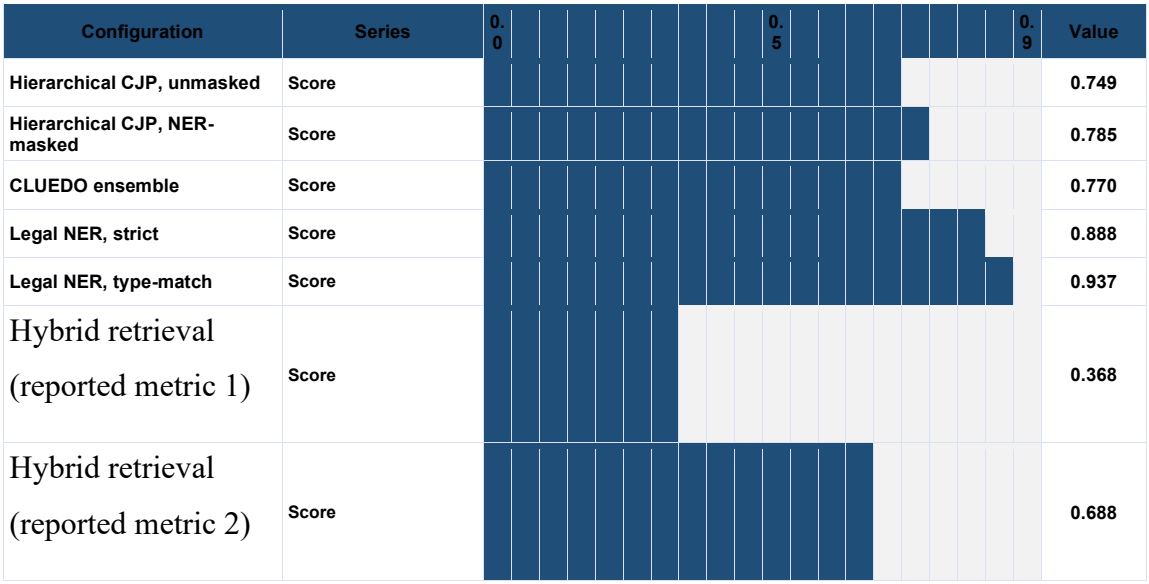


Fig. 16. Selected legal-AI task metrics. Source: Benedetto (2025) and Benedetto, Cagliero and Tarasconi (2025); metrics are task-specific and are not pooled. [11,12]

Figure 16 groups heterogeneous legal-NLP scores only to demonstrate task and metric dependence; the scores are not treated as directly commensurable or combined into CARI.

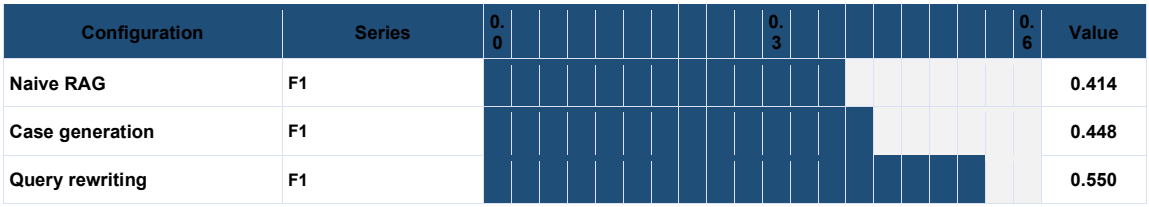


Fig. 17. Legal RAG configuration comparison. Source: Benedetto (2025) and Benedetto, Cagliero and Tarasconi (2025). [11,12]

Figure 17 shows the reported F1 improvement from naive RAG to case generation and query rewriting, supporting the narrower claim that retrieval design can materially change task performance.

Table F1. Legal RAG configuration results reported by Benedetto [11,12]

Configuration	Metric	Reported value	Interpretive boundary
Naive RAG	F1	0.414	Task-specific retrieval/reasoning metric; not legal deployment reliability
Case generation	F1	0.448	Task-specific metric; not transferable to unrelated legal tasks
Query rewriting	F1	0.550	Improves the reported task relative to naive RAG; does not imply equivalent gain in legal correctness

Table F1 isolates the comparable RAG configurations and shows the absolute and relative improvement associated with query rewriting while preserving the task-specific interpretive boundary.

Table F2. Exact values displayed in Figure 16

Configuration / task label	Reported score	Interpretive boundary
Hierarchical CJP, unmasked	0.749	Task-specific score; do not pool across tasks
Hierarchical CJP, NER-masked	0.785	Task-specific score; do not pool across tasks
CLUEDO ensemble	0.770	Task-specific score; do not pool across tasks
Legal NER, strict	0.888	NER metric; not equivalent to RAG F1
Legal NER, type-match	0.937	NER metric; not equivalent to RAG F1
Hybrid retrieval, reported metric 1	0.368	Separate reported retrieval metric/condition
Hybrid retrieval, reported metric 2	0.688	Separate reported retrieval metric/condition

Table F2 makes the values underlying Figure 16 explicit and also records why the heterogeneous scores are not interpreted as one common performance scale.

Appendix G. Randomized legal-work extraction

Participant-level data were not reanalysed and no retrospective multiple-comparison correction is imposed. The appendix therefore reports task-level estimates and their published statistical interpretation without upgrading uncertain effects.

Task	Quality_change	p_value	Sign
Complaint drafting	+0.17	0.0862	Positive
Contract drafting	+0.24	0.0060	Positive
Employee handbook	+0.07	0.3532	Positive
Client memorandum	-0.07	0.5980	Negative

Fig. 18. Task-level quality effects in randomized GPT-4-assisted legal work. Source: Choi, Monahan and Schwarcz (2024). [3]

Figure 18 shows positive quality estimates for three randomized legal tasks and a negative estimate for the client memorandum, making the task-specific nature of quality effects visible.

Table G1. Interpretation of task-level randomized evidence

Task	Time reduction	Quality change	Statistical interpretation
Complaint drafting	24.1%	+0.17	Uncertain
Contract drafting	32.1%	+0.24	Supported
Employee handbook	21.1%	+0.07	Uncertain
Client memorandum	11.8%	-0.07	Uncertain

Table G1 aligns each task's time reduction with its quality estimate and significance status, making clear that faster completion did not imply a uniform quality improvement.

Appendix H. Training intervention analysis

The training study is retained as causal evidence about intervention design within one examination context, while its preprint status and setting limit generalization. It supports testing reviewer readiness rather than assuming that professional status alone guarantees effective oversight.

Table H1. Training-study extraction

Comparison	Estimate	Reported p-value	CARI implication
Adoption: trained vs untrained	+15 percentage points	0.044	Access is not equivalent to use
Grade: trained vs untrained	+0.27 grade points	0.027	Training can increase review/workflow readiness
Applicable-rule accuracy	Improved	0.014	Review competence should be tested on legal substance

Table H1 records the adoption, grade and legal-rule outcomes that motivate treating training and demonstrated competence as measurable parts of reviewer readiness.

Appendix I. Professional adoption, use and agentic AI

The Thomson Reuters survey describes organizational exposure and reported use, not causal effectiveness. Figures 19-22 separate tool type, use frequency, legal use case and agentic adoption so that different forms of deployment are not collapsed into one adoption statistic.

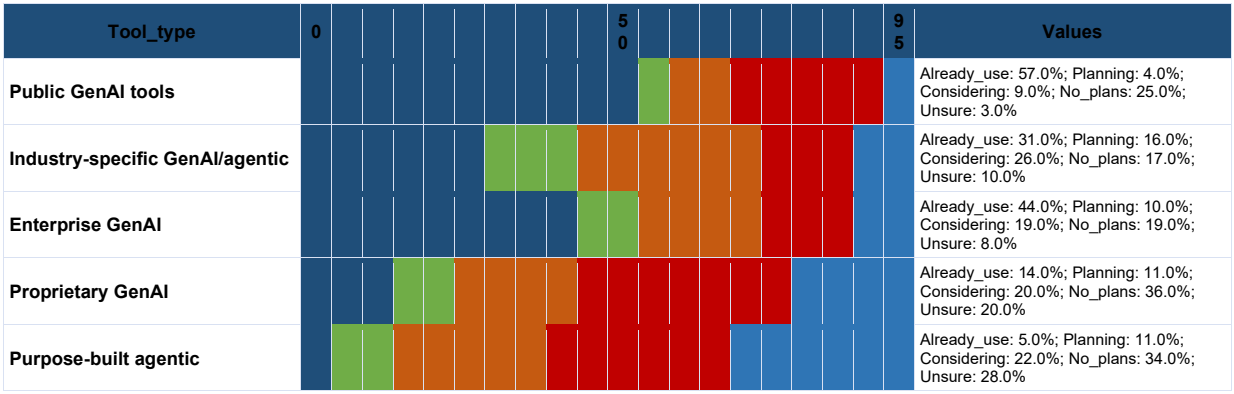


Fig. 19. Individual use and adoption intentions by AI tool type. Source: Thomson Reuters Institute (2026). [7]

Figure 19 distinguishes current use, planned use and consideration across public, enterprise, industry-specific, proprietary and purpose-built agentic tools, describing adoption pathways rather than effectiveness.

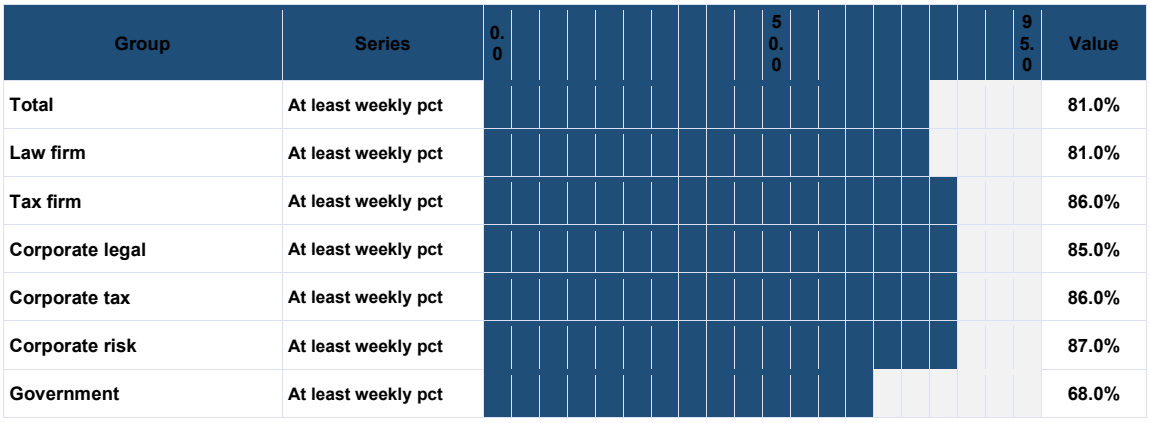


Fig. 20. At-least-weekly GenAI use among current users. Source: Thomson Reuters Institute (2026). [7]

Figure 20 shows that most current users report at-least-weekly GenAI use, indicating that exposure is embedded in routine professional workflows rather than limited to occasional experimentation.

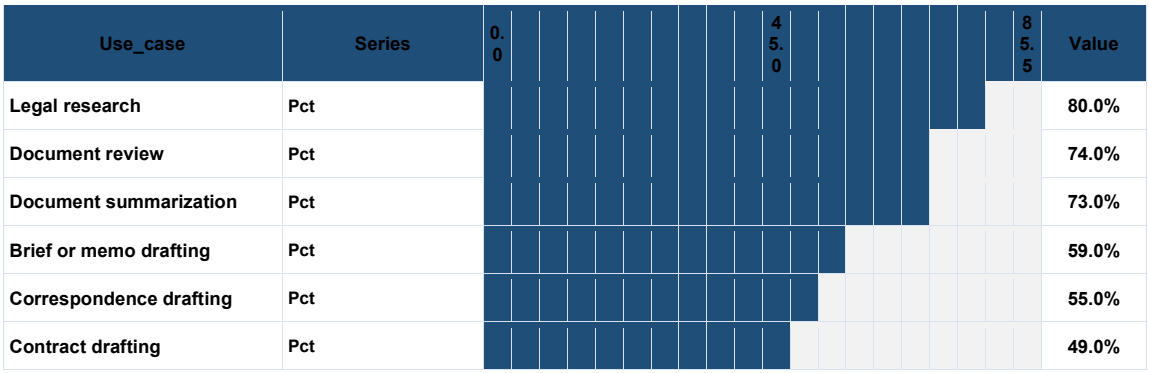


Fig. 21. Top reported legal GenAI use cases. Source: Thomson Reuters Institute (2026). [7]

Figure 21 identifies research, review and summarization as the leading reported legal use cases, which are precisely the tasks for which source verification and escalation controls are especially relevant.

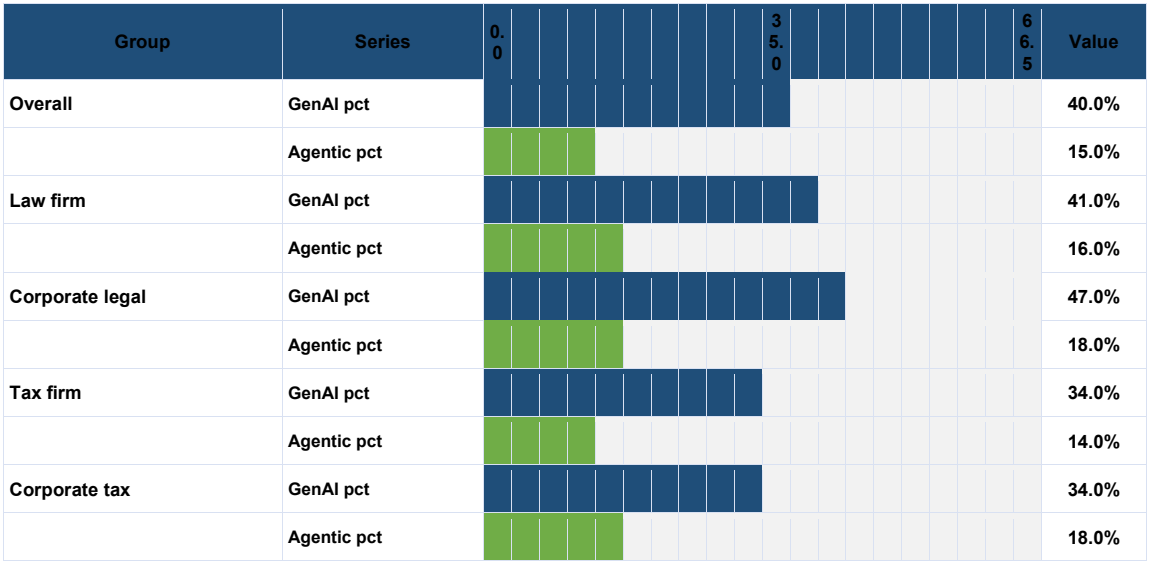


Fig. 22. Current GenAI and agentic-AI organizational use. Source: Thomson Reuters Institute (2026). [7]

Figure 22 separates current GenAI use from agentic-AI use across organizational groups and therefore avoids treating all forms of automation as one adoption category.

Table I1. Professional adoption and measurement indicators

Indicator	Reported value	Interpretation
Organizations already using GenAI	40%	Up from 22% in the prior survey
Current users using GenAI at least weekly	82%	Use is embedded in regular work
Organizations collecting AI ROI metrics	18%	Measurement substantially trails adoption
Organizations already using agentic AI	15%	Early adoption with 53% planning or considering

Indicator	Reported value	Interpretation
Expect agentic AI central by 2030	77%	Anticipated workflow dependence
Corporate respondents wanting outside firms to use AI	Approximately two-thirds	Demand is not matched by clear mandates

Table I1 supplies the headline adoption, weekly-use and measurement values that anchor the professional-use figures while emphasizing that usage prevalence is not an effectiveness estimate.

Appendix J. Consumer expectations and protection gap

Consumer evidence is used to identify expectations and perceived protection needs rather than to determine technical reliability. Net balances complement the full better/worse percentages reported in the main text.

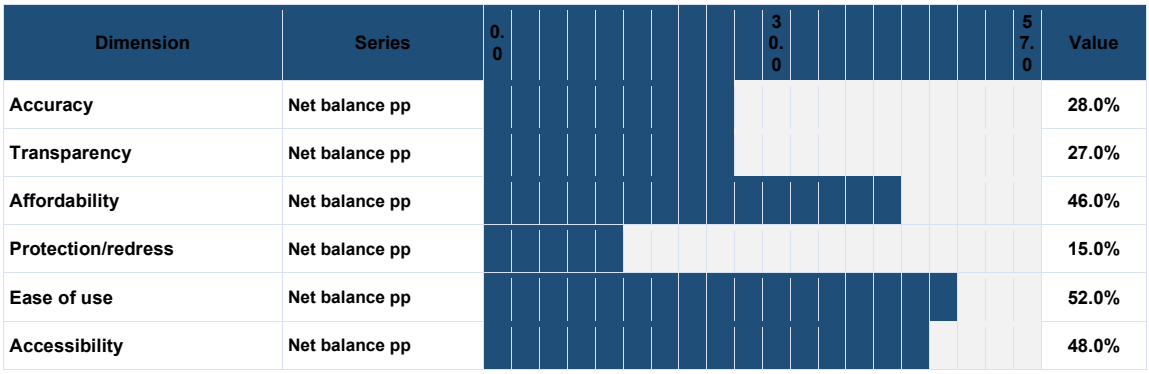


Fig. 23. Net consumer optimism by dimension. Source: Legal Services Board (2026). [8]

Figure 23 converts the consumer survey into net balances and shows that optimism is strongest for ease, accessibility and affordability and weaker for protection/redress.

Table J1. Additional Legal Services Board findings

Finding	Value	Interpretation
Expect easier legal services	70%	Strongest anticipated benefit
Expect better accessibility	66%	Access-to-justice potential
Expect greater affordability	64%	Cost benefit expected
Standards identified as non-binding guidance	58%	Enforcement gap in landscape review
Favour mandatory regulation	57%	Preference for enforceable baseline

Finding	Value	Interpretation
Adults with legal problem receiving no professional support	Approximately 32%	Potential demand for direct-to-consumer tools

Table J1 adds the most salient access-oriented findings and clarifies that expected accessibility and affordability gains coexist with a demand for stronger protection and redress.

Appendix K. Role-differentiated trust, US firm-size adoption and implementation concerns

The Walker case study is exploratory and purposive, whereas the ABA evidence is descriptive survey evidence. They are retained together because both address how organizational role, scale and perceived risk shape the conditions under which legal professionals rely on AI.

Role	Series	0.0 50.0 95.0	Value
Consumer	Difficulty defining AI pct		60.0%
	AI opaque pct		60.0%
Producer	Difficulty defining AI pct		0.0%
	AI opaque pct		20.0%
Liminal	Difficulty defining AI pct		83.3%
	AI opaque pct		66.7%

Fig. 24. Role-differentiated understanding and perceived opacity. Source: Walker (2024), purposive two-firm case study, n=21. [10]

Figure 24 shows marked role differences in difficulty defining AI and perceived opacity; the small purposive sample limits generalization but supports role-calibrated assurance communication.

Firm_size	Series	0.0 27.5 52.2	Value
All respondents	AI use pct		30.2%
500+ lawyers	AI use pct		47.8%
10-49 lawyers	AI use pct		29.5%
2-9 lawyers	AI use pct		24.1%
Solo	AI use pct		17.7%

Fig. 25. US office AI use by firm size. Source: American Bar Association Legal Technology Resource Center (2025). [9]

Figure 25 shows higher reported office AI use in larger US firms, a descriptive exposure pattern that does not establish a causal firm-size effect.

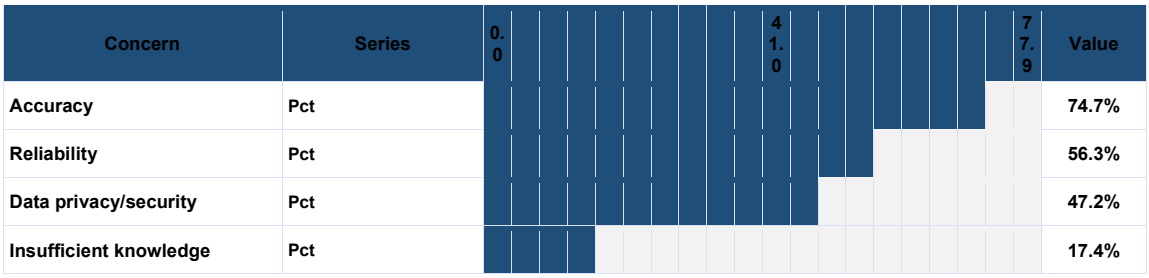


Fig. 26. Reported US AI implementation concerns. Source: American Bar Association Legal Technology Resource Center (2025). [9]

Figure 26 shows that accuracy, reliability and privacy/security dominate reported implementation concerns, aligning organizational concerns with the paper's reliability and governance dimensions.

Table K1. Role-level case-study indicators

Role	Difficulty defining AI	Perceived opacity	n
Consumer	60.0%	60.0%	10
Producer	0.0%	20.0%	5
Liminal	83.3%	66.7%	6

Table K1 provides the small-sample role counts behind Figure 24 and should be interpreted as exploratory evidence about differentiated understanding, not population prevalence.

Table K2. Reported US office AI use by firm size

Firm-size group	Office AI use
All respondents	30.2%
500+ lawyers	47.8%
10-49 lawyers	29.5%
2-9 lawyers	24.1%
Solo	17.7%

Table K2 provides the exact firm-size percentages behind Figure 25 and supports only a descriptive adoption gradient.

Table K3. Reported US AI implementation concerns

Implementation concern	Respondents
Accuracy	74.7%

Implementation concern	Respondents
Reliability	56.3%
Data privacy/security	47.2%
Insufficient knowledge	17.4%

Table K3 lists the concern frequencies behind Figure 26, showing that accuracy and reliability are the dominant reported implementation issues.

Appendix L. Jurisdictional coding rubric, descriptive matrix and numeric sensitivity view

The primary comparative result is the descriptive legal-mechanism matrix. The original 0/0.5/1 coding and normalized coverage figure are restored only to make the historical coding procedure auditable. They are not used as CARI G, do not measure enforcement, institutional capacity or outcomes, and must not be read as a jurisdictional league table.

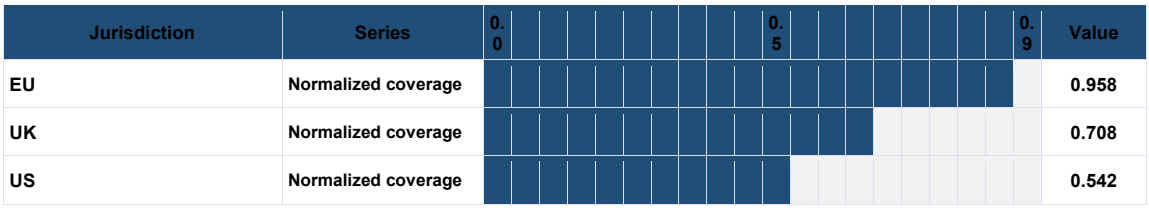


Fig. 27. Normalized formal-safeguard coverage under the transparent equal-weight coding sensitivity view. Source: Author coding of official instruments. This is not CARI G and does not measure enforcement quality or outcomes. [15-24]

Figure 27 summarizes the original numeric legal coding as a transparent sensitivity view only; the main analysis relies on the descriptive safeguard matrix and does not use these normalized values as organizational G.

Table L1. Original numeric safeguard coding matrix used for the descriptive sensitivity view

Safeguard	EU	UK	US
Cross-sector binding AI law	1.0	0.0	0.0
Justice-sector risk classification	1.0	0.5	0.0
Ex ante risk management	1.0	0.5	0.5
Data governance	1.0	1.0	0.5
Technical documentation	1.0	0.5	0.5

Safeguard	EU	UK	US
Logging/record keeping	1.0	0.5	0.5
Human oversight	1.0	1.0	1.0
Accuracy/robustness/cybersecurity	1.0	0.5	0.5
Transparency to affected persons	1.0	1.0	0.5
Contestability/redress	0.5	1.0	0.5
Professional ethics guidance	1.0	1.0	1.0
Post-deployment monitoring/enforcement	1.0	1.0	1.0

Table L1 shows exactly how the original equal-weight numeric visualization was produced; its values are retained for reproducibility but are not organizational governance measurements.

Table L2. Primary descriptive safeguard coverage matrix

Safeguard	EU	UK	US
Cross-sector binding AI law	Present	No general mechanism identified	No general mechanism identified
Justice-sector risk classification	Present	Partial/sectoral	No general mechanism identified
Ex ante risk management	Present	Partial/sectoral	Partial/sectoral
Data governance	Present	Present	Partial/sectoral
Technical documentation	Present	Partial/sectoral	Partial/sectoral
Logging/record keeping	Present	Partial/sectoral	Partial/sectoral
Human oversight	Present	Present	Present
Accuracy/robustness/cybersecurity	Present	Partial/sectoral	Partial/sectoral
Transparency to affected persons	Present	Present	Partial/sectoral
Contestability/redress	Partial/sectoral	Present	Partial/sectoral

Safeguard	EU	UK	US
Professional ethics guidance	Present	Present	Present
Post-deployment monitoring/enforcement	Present	Present	Present

Table L2 is the preferred legal comparison because it reports whether a general, partial/sectoral or no comparable mechanism was identified without implying quantitative equivalence among legal safeguards.

Appendix M. CARI formal properties, decision algorithm and counterfactual sensitivity stress test

CARI is a prospective non-compensatory eligibility rule. The formal objects and decision algorithm below are part of the proposed framework. The candidate floors, scenario scores, pass-probability ranges and bottleneck figure that follow are retained solely as counterfactual stress tests of the mechanics; they require future empirical calibration and are not safe harbours.

Table M0. Formal objects and prospective validation targets

Formal object	Definition	Observable interpretation	Prospective test
CARI gate	CARI_s=1 only if R_L≥r_s, G≥g_s and H≥h_s	Non-compensatory eligibility after local measurement	Compare with observed material errors, corrections and incidents under prospectively defined outcomes
Assurance margin	AM_s=min(R_L-r_s,G-g_s,H-h_s)	Signed distance to the weakest measured floor	Test whether larger margins predict lower material-error or correction rates
Measurement sufficiency rule	No live CARI decision if R_L, G or H is unmeasured	Prevents public benchmarks or policy proxies from becoming deployment certification	Audit whether missing local evidence predicts inappropriate reliance decisions
Bottleneck attribution	Identity of the minimum term in AM_s	Weakest measured component requiring correction	Compare attributed bottlenecks with later audit findings and reviewer failures
Expiry rule	Approval invalid after predefined material model, corpus, workflow or legal change	Version-bound validity	Measure performance drift before and after change triggers

Table M0 defines the CARI gate, assurance margin, measurement-sufficiency rule, bottleneck attribution and expiry rule, and states the prospective tests required before operational validation.

Table M1. CARI decision algorithm

Step	Operation	Minimum auditable evidence
1	Define use case and stakes tier	Users, consequence, reversibility and prohibited autonomous actions
2	Freeze configuration	Model/provider version, corpus, prompt workflow, jurisdiction and date
3	Estimate conservative reliability	Representative test set, error taxonomy and lower bound
4	Audit governance	Logs, access controls, data rules, incident and redress process
5	Test review readiness	Seeded errors, source verification, overrides and time budget
6	Apply predeclared floors and assurance margin	CARI pass/fail only after R_L, G and H are locally measured; binding bottleneck recorded
7	Record expiry and retest trigger	Approval owner, sunset date and change criteria

Table M1 converts the conceptual framework into an auditable sequence from use-case definition and configuration freezing to local testing, decision, monitoring and expiry.

Table M2. Candidate stakes-dependent floors retained for counterfactual stress testing only

Tier	Illustrative legal use	r_s	g_s	h_s	Additional condition
Low	Internal drafting, triage, brainstorming	0.6	0.6	0.7	Source verification before external use
Medium	Client advice, contracts, settlement analysis	0.75	0.75	0.85	Named lawyer approval and logged overrides
High	Court filing or liberty/property decision	0.9	0.9	0.95	Independent verification, redress and incident reporting

Table M2 records the original illustrative low, medium and high floors. These values are hypotheses for future calibration, not validated legal or statistical thresholds.

Table M3. Deterministic illustrative CARI scenarios under declared assumptions

Scenario	R_L	G	H	Legacy geometric summary (not used in CARI eligibility)	Eligibility
Open general chatbot	0.420	0.350	0.500	0.419	No tier
Commercial legal RAG	0.580	0.600	0.650	0.609	No tier
Domain-specific supervised model	0.785	0.750	0.800	0.778	Low only
Audited RAG/reasoning + mandatory review	0.800	0.900	0.900	0.865	Low + medium

Table M3 shows how the non-compensatory gate behaves when hypothetical R_L, G and H values are supplied. The retained geometric column is a legacy descriptive summary from the earlier index formulation and plays no role in the current binary CARI eligibility rule.

Table M4. Counterfactual CARI pass-probability sensitivity ranges

Profile	Low-tier gate-crossing frequency	Medium-tier gate-crossing frequency	High-tier gate-crossing frequency
Profile A (65%)	92.6% [91.7–92.6]	0.1% [0.0–0.1]	0.0% [0.0–0.0]
Profile B (42%)	0.0% [0.0–0.0]	0.0% [0.0–0.0]	0.0% [0.0–0.0]
Profile C (20%)	0.0% [0.0–0.0]	0.0% [0.0–0.0]	0.0% [0.0–0.0]
Profile D (49%)	0.1% [0.1–0.1]	0.0% [0.0–0.0]	0.0% [0.0–0.0]

Table M4 retains the original simulation frequencies for transparency but reports them as anonymized benchmark profiles. They depend on benchmark accuracy used as a legacy reliability proxy, assumed G/H distributions and candidate floors; they are sensitivity diagnostics, not empirical product approval probabilities.

Profile	Series	0.0 50.0 95.0	Value
Profile C (20%)	Reliability fail		100.0%
	Governance fail		0.1%
	Review fail		10.6%
Profile D (49%)	Reliability fail		100.0%
	Governance fail		0.1%
	Review fail		10.9%
Profile A (65%)	Reliability fail		99.9%
	Governance fail		0.1%
	Review fail		11.3%
Profile B (42%)	Reliability fail		100.0%
	Governance fail		0.1%
	Review fail		11.1%

Fig. 28. Illustrative failure-component frequencies for four anonymized benchmark profiles under declared candidate floors and hypothetical governance/reviewer inputs. Source: Author counterfactual sensitivity analysis. Failure components overlap and are not empirical approval probabilities.

Figure 28 decomposes the counterfactual exercise into overlapping reliability, governance and reviewer failure components. Its purpose is to illustrate bottleneck logic under assumptions, not to estimate the safety of a product or jurisdiction.

Appendix N. Residual-error and normalized value sensitivity illustrations

The residual-error equation is a didactic control model rather than a live malpractice estimate. For completeness, the earlier normalized Net Accountable Value (NAV) illustration is retained as a legacy exploratory sensitivity analysis. NAV is not part of CARI, its benefit and governance-cost inputs are not empirical monetary estimates, and no economic or policy conclusion is drawn from it. NAV is defined as B_gross - C_gov - E_res c_e, with the break-even error cost c_e*=(B_gross-C_gov)/E_res under the declared normalized scenario.

Baseline_label	0	25	50	75	100
17%	1700	1360	1020	680	340
33%	3300	2640	1980	1320	660
43%	4300	3440	2580	1720	860

Fig. 29. Residual errors as review coverage increases. Source: Illustrative equation E_res=Np_e(1-hd), N=10,000 and d=0.80.

Figure 29 demonstrates the mechanical effect of review coverage in the residual-error equation: incomplete coverage leaves a substantial residual even when reviewers are reasonably competent.

Baseline_label	50	60	70	80	90	100
17%	892	731	570	408	246	—
33%	1732	1419	1105	792	478	—
43%	2257	1849	1440	1032	623	—

Fig. 30. Residual errors as detection competence improves. Source: Illustrative equation E_res=Np_e(1-hd), N=10,000 and h=0.95.

Figure 30 demonstrates the complementary effect of detection competence: high review coverage alone cannot eliminate residual errors when reviewers fail to identify material defects.

Baseline_label	70	80	90	95
17%	2.46	3.43	5.68	8.45
33%	1.27	1.77	2.93	4.35
43%	0.97	1.36	2.25	3.34

Fig. 31. Legacy normalized NAV break-even sensitivity. Source: Author didactic scenario with N=10,000, 19% gross benefit and governance cost=500 normalized units. This is an illustrative sensitivity calculation, not an empirical monetary cost-benefit model.

Figure 31 shows only the break-even mechanics of a normalized illustrative value model; because the inputs are not empirical monetary estimates, the figure is not used to claim economic value.

Table N1. Selected residual-error scenarios

Baseline error	Review coverage	Detection	Residual errors/10,000
17%	95%	80%	408
17%	100%	95%	85
33%	95%	80%	792
43%	95%	80%	1,032

Table N1 provides exact values for representative combinations of baseline error, review coverage and conditional detection, making the assumptions behind Figures 29-30 explicit.

Appendix O. Versioning, documentation cadence and source-to-claim reproducibility

A reliance decision is configuration-specific and should expire after a material change to the model, provider version, retrieval corpus, prompt workflow, legal environment or review process. The documentation-cadence figure is illustrative rather than exhaustive and is used to show why version identifiers and dates belong in the assurance record.

Month	Provider	Documented_updates
2026-02-01	Anthropic	2
2026-02-01	Google DeepMind	3
2026-03-01	OpenAI	1
2026-03-01	Google DeepMind	1
2026-04-01	OpenAI	2
2026-04-01	Anthropic	2
2026-04-01	Google DeepMind	1
2026-05-01	OpenAI	1
2026-05-01	Anthropic	1
2026-05-01	Google DeepMind	2
2026-06-01	OpenAI	2
2026-06-01	Anthropic	2
2026-06-01	Google DeepMind	2
2026-07-01	OpenAI	2
2026-07-01	Anthropic	1
2026-07-01	Google DeepMind	2

Fig. 32. Selected official model/system-card documentation cadence. Source: Official OpenAI, Anthropic and Google DeepMind records, accessed by the study data cut-off. [25-27]

Figure 32 illustrates how frequently public model documentation can change over time, supporting the requirement that a reliance decision attach to a frozen configuration and expire after material change.

Table O1. Source-to-claim and reproduction register

Claim/data object	Primary source	Reproduction file
Legal-research outcome percentages	Magesh et al. (2025)	benchmark_outcomes.csv
Wilson intervals and risk ratios	Author calculation	wilson_rounding_envelope.csv; risk_ratios_descriptive.csv; python/analysis_reference.py
SimpleQA composition	OpenAI GPT-5 System Card	openai_simpleqa.csv
Task time and quality	Choi et al. (2024)	choi_rct_tasks.csv
Training effects	Chen and Bao (2026)	chen_training.csv
Professional adoption and use	Thomson Reuters Institute (2026)	tri_*.csv
Consumer expectations	Legal Services Board (2026)	lsb_consumer_expectations.csv
Safeguard coding	Official EU/UK/US instruments	jurisdiction_safeguards.csv
CARI simulation	Author analysis	counterfactual_gate_crossing_summary.csv; cari_bottlenecks_summary.csv

Claim/data object	Primary source	Reproduction file
Residual error and NAV	Author analysis	residual_error_coverage.csv; residual_error_detection.csv; nav_sensitivity.csv; python/analysis_reference.py
Wilson power, precision decay & risk differences (Figs. 33–35)	Eq. (1) & Magesh et al. (2025)	OriginPro_Master_Data_47_Figures.xlsx (Fig33–Fig35)
Selective prediction, RCT quadrant & training CIs (Figs. 36–38)	OpenAI (2025); Choi et al. (2024); Chen & Bao (2026)	OriginPro_Master_Data_47_Figures.xlsx (Fig36–Fig38)
Governance gap, agentic error & consumer tension (Figs. 39–41)	TRI (2026); LSB (2026); exact 1−p^k model	OriginPro_Master_Data_47_Figures.xlsx (Fig39–Fig41)
Dirichlet weight sensitivity, 2D contour & fatigue model (Figs. 42–44)	Table L1; Eq. (4) & Eq. (4b)	OriginPro_Master_Data_47_Figures.xlsx (Fig42–Fig44)
Bayesian R_L updating, CARI tornado & drift survival (Figs. 45–47)	Beta-Binomial model; Eq. (3); Appendix O records	OriginPro_Master_Data_47_Figures.xlsx (Fig45–Fig47)

Table O1 links each principal claim or derived object to its primary source and reproduction file, providing the audit trail needed to reproduce the quantitative and legal analyses.

Declarations

Funding: No external funding was received for this study.

Competing interests: The author declares no competing interests.

Ethics: The study analyses published and lawfully accessible secondary evidence and did not collect new human-participant data.

Data and code availability: A complete reproducibility package accompanies this manuscript, containing the OriginPro/Excel master workbook (OriginPro_Master_Data_47_Figures.xlsx), the OriginPro automation script (Generate_All_Figures_OriginPro.py), and exact statistical verification outputs indexed in Appendix O. Every figure in this manuscript (Figures 1–49) is retained as a native, editable Word object with zero rasterized images so that all displayed values remain directly inspectable and editable.

Use of generative AI: Generative-AI tools were used for language editing, manuscript organization, figure/table quality control and consistency checking. AI-generated text was not treated as legal authority or empirical evidence. Factual and legal claims are supported by cited

sources or explicitly identified as author assumptions or sensitivity scenarios. Source verification, legal interpretation and final responsibility for the submitted manuscript remain with the author.

References

[1] Magesh, Varun, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. "Hallucination-free? Assessing the reliability of leading AI legal research tools." Journal of empirical legal studies 22, no. 2 (2025): 216-242.

[2] Dahl, Matthew, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. "Large legal fictions: Profiling legal hallucinations in large language models." Journal of Legal Analysis 16, no. 1 (2024): 64-93.

[3] Choi, Jonathan H., Amy B. Monahan, and Daniel Schwarcz. "Lawyering in the age of artificial intelligence." Minn. L. Rev. 109 (2024): 147.

[4] Nielsen, Aileen, Stavroula Skylaki, Milda Norkute, and Alexander Stremitzer. "Building a better lawyer: Experimental evidence that artificial intelligence can increase legal work efficiency." Journal of Empirical Legal Studies 21, no. 4 (2024): 979-1022.

[5] Schwarcz, Daniel, Sam Manning, J. J. Prescott, Patrick Barry, David R. Cleveland, and Beverly Rich. "AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice." Journal of Law & Empirical Analysis 3, no. 1 (2026): 220-250.

[6] Chen, Benjamin M., and Hong Bao. "Training for technology: Adoption and productive use of generative AI in legal analysis." arXiv preprint arXiv:2603.04982 (2026).

[7] Thomson Reuters Institute (2026) 2026 AI in Professional Services Report. Survey of 1,514 AI-familiar professionals in 27 countries. https://www.thomsonreuters.com/en/reports/2026-ai-in-professional-services-report

[8] Legal Services Board (2026) AI in legal services: consumer expectations and existing standards. Published 10 June 2026. https://legalservicesboard.org.uk/research-2/ai-in-legal-services-consumer-expectations-and-existing-standards%20(accessed%201%20August%202026.

[9] American Bar Association Legal Technology Resource Center (2025) 2024 Artificial Intelligence TechReport. https://www.americanbar.org/groups/law_practice/resources/tech-report/2024/2024-artificial-intelligence-techreport/

[10] Walker, Edward Alastair. "Questioning producers and accepting consumers: responses to artificial intelligence amongst legal professionals." PhD diss., University of Sheffield, 2024.

[11] Benedetto, Irene. "Enhancing Legal Document Processing through Natural Language Understanding and Generation techniques." (2025).

[12] Benedetto I, Cagliero L, Tarasconi F (2025) LR2: A Legal RAG for Legal Reasoning over Cases. Frontiers in Artificial Intelligence and Applications. https://doi.org/10.3233/FAIA251614

[13] Afridi, Hina, Habib Ullah, Sultan Daud Khan, and Mohib Ullah. "From GPT-3 to GPT-5: Mapping their capabilities, scope, limitations, and consequences." arXiv preprint arXiv:2604.10332 (2026).

[14] Wooldridge, Michael, Peter Asaro, Lucie Flek, Iryna Gurevych, Sara Hooker, Gina Neff, Peter Norvig et al. "AI goes mainstream: The scaling era and the rise of generative AI. The One Hundred Year Study on Artificial Intelligence (AI100). 2026 study panel report." (2026).

[15] Lognoul, Michael. "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act–AI Act)." Revue du Droit des Technologies de l'information 3-4 (2025): 145-189.

[16] UK, GOV. "A pro-innovation approach to AI regulation: government response." Online verfügbar unter https://www. gov. uk/government/consultations/ai-regulation-a-pro-innovationapproach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-governmentresponse, zuletzt aktualisiert am 24 (2024): 2024.

[17] United Kingdom (2025) Data (Use and Access) Act 2025 and explanatory notes on automated decision-making safeguards. https://www.legislation.gov.uk/ukpga/2025/18/notes/division/10/index.htm

[18] Manuel, Marcus, Clare Manuel, and Yoko Yoshihara. "AI and the potential for universal affordable access to front-line justice." (2026).

[19] Solicitors Regulation Authority (2026) Risk Outlook report: AI and legal services. Updated 23 April 2026. https://www.sra.org.uk/sra/research-publications/risk-outlook-paper-innovation-competitive-landscape/

[20] American Bar Association. "Formal Opinion 512: Generative artificial intelligence tools." ABA Standing Committee on Ethics and Professional Responsibility (2024).

[21] AI, NIST. "Artificial intelligence risk management framework (AI RMF 1.0)." URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai 1, no. 0 (2023): 0-1.

[22] AI, NIST. "Artificial intelligence risk management framework: Generative artificial intelligence profile." NIST Trustworthy and Responsible AI Gaithersburg, MD, USA (2024).

[23] Alibašić, Haris. "Governing Agentic AI in the Administrative State: Human Oversight, Cybersecurity Accountability, and Risk in Autonomous Digital Systems." Future Internet 18, no. 10 (2026): 528.

[24] Federal Trade Commission. "FTC announces crackdown on deceptive AI claims and schemes (Operation AI Comply)." September. Accessed July 22 (2024): 2026.

[25] Killekar, MS Aditya, MD Robert JH Miller, PhD Damini Dey, MS Paul B. Kavanagh, MD Jamieson M. Bourque, MD Lawrence M. Phillips, MD Panithaya Chareonthaitawee, and PhD Piotr J. Slomka. "Retrieval-augmented claude opus 4.7 and GPT-5.5 surpass human performance on the nuclear cardiology board preparation exam (and claude drafts a paper about it)." Journal of Nuclear Cardiology (2026): 102946.

[26] Fletcher, Gordon, and Saomai Vu Khan. "Pathways to AGI." arXiv preprint arXiv:2605.06029 (2026).

[27] Goworek, Roksana. "Open Boxing of AI Systems: Activation-Based Interpretability for AI Safety." (2026).

[28] Bertolini, Andrea, and Shabahang Arian. "Automated care-taking and the constitutional rights of the patient in an aging population." In Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice, pp. 297-321. The Hague: TMC Asser Press, 2022.

[29] Surden, Harry. "Artificial intelligence and law: An overview." Ga. St. UL Rev. 35 (2018): 1305.

[30] Ashley, Kevin D. "Artificial intelligence and legal analytics." New Tools for Law (2017).

[31] Guha, Neel, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters et al. "Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models." Advances in neural information processing systems 36 (2023): 44123-44279.

[32] Chien, Colleen V., and Miriam Kim. "Generative AI and legal aid: Results from a field study and 100 use cases to bridge the access to justice gap." Loy. LAL Rev. 57 (2024): 903.

[33] Legal Services Board (2024) The legal needs of individuals in England and Wales: findings from the 2023 Individual Legal Needs Survey. https://legalservicesboard.org.uk/research-2/individual-legal-needs-survey%20(accessed%201%20August%202026

[34] Terzidou, Kalliopi. "Generative AI systems in legal practice offering quality legal services while upholding legal ethics." International Journal of Law in Context 21, no. 3 (2025): 431-452.

[35] Sourdin, Tania, and Bin Li. "AI Governance: A Comparative Approach." German Law Journal 26, no. 7 (2025): 1099-1116.

[36] Weinstein, Stuart. "Public-facing generative-AI legal guidance systems: applied legal-ethics, reliability and governance limits–a multi-case study." International Journal of Law and Management (2026): 1-27.

[37] Kourabas, Steve, and Cheng-Yun Tsang. "The board monitoring function: Artificial intelligence in the era of heightened accountability." The Chinese Journal of Comparative Law 13 (2025): cxaf013.

[38] Gaube, Susanne, Markus Langer, Tim Miller, Kevin Baum, Raimund Dachselt, Anna Maria Feit, Ujwal Gadiraju et al. "Keeping an Eye on AI: A Framework for Effective Human Oversight of AI Systems." arXiv preprint arXiv:2605.16278 (2026).

[39] Laux, Johann. "Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act." AI & society 39, no. 6 (2024): 2853-2866.

[40] Zhu, Liming, Qinghua Lu, Ming Ding, Sung Une Lee, and Chen Wang. "Designing meaningful human oversight in AI." AI and Ethics 6, no. 3 (2026): 286.