International Journa of Advanced AI Applications

Volume2,Issue9,Sep.2026

Online ISSN 3104-9338

Print ISSN 3104-932X

Hong Kong Dawn Clarity Press Limited

http://www.dawnclarity.press/index.php/ijaaa

TABLE OF CONTENTS

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

Safety and Accuracy of AI-Generated Return-to-Sport Recommendations After ACL Reconstruction: A Content Analysis Against the Bern Consensus

Tongtong Dai[1], Yiming Wang*

Zhejiang Normal University, Jinhua 321004, China

Received: August 19, 2026

Revised: August 19, 2026

Accepted: August 20, 2026

Published online: August 26, 2026

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

* Corresponding Author: Yiming Wang (wangyiming9616@outlook.com)

Abstract. Large language models (LLMs) are increasingly consulted by athletes for return-to-sport (RTS) guidance after anterior cruciate ligament (ACL) reconstruction, yet their alignment with clinical consensus remains unknown. This study evaluated the accuracy, completeness, and safety of LLM-generated RTS recommendations against the 2016 Bern Consensus criteria. Fifteen clinical vignettes (five standard, five borderline, five high-risk) were each queried three times across three LLMs (DeepSeek-V3, Grok-4-Fast, GPT-5-nano), yielding 135 responses scored on a Bern Consensus – derived six-dimension framework, supplemented by a 3-level Likert sensitivity analysis for psychological readiness and sport-specific skills. All models achieved perfect scores for time factor, risk – benefit discussion, and safety warnings. Sport-specific skills were nearly absent (Likert scores below 0.07), and psychological readiness varied by model (DeepSeek 0.42, GPT 0.23, Grok 0.00). In high-risk cases, data hallucination and comorbidity omission reduced mean total scores from 4.36 to 3.69, and mixed-effects modelling confirmed significant case-type effects ($p=0.001$). Current LLMs demonstrate adequate baseline ACL knowledge but fail in sport-specific precision; AI-generated RTS advice should therefore serve only as a preliminary draft requiring mandatory individualized clinical validation.

Keywords: *Large Language Models; Sports Physical Therapy; Clinical Decision-making; Orthopedic Rehabilitation; Psychological Readiness*

1. Introduction

Anterior cruciate ligament (ACL) reconstruction is among the most common orthopedic

procedures in young athletes, with an annual incidence exceeding 250,000 cases in the United States alone [1,2]. The postoperative endpoint is not structural healing but safe return to sport (RTS)—a multidimensional decision integrating biological recovery, functional performance, psychological readiness, and sport-specific skill acquisition [3]. Yet surgical success does not guarantee athletic success: a landmark systematic review and meta-analysis of 48 studies including 5,770 participants found that although 82% of patients returned to some form of sports participation, only 63% regained their preinjury level of participation and only 44% returned to competitive sport, with fear of re-injury the most frequently cited reason for reducing or ceasing participation [4].

To standardize this decision, the 2016 Bern Consensus Statement established that RTS should be criterion-based, requiring objective functional thresholds (>90% limb symmetry index), psychological confidence, and sport-specific movement competency before full competition is permitted [3]. This framework has become the reference standard for RTS decision-making in sports physical therapy practice worldwide. Subsequent consensus work, such as the Panther Symposium, further refined RTS terminology along a continuum extending from return to participation, through return to sport, to return to performance [5]. In parallel, dedicated instruments such as the ACL Return to Sport after Injury (ACL-RSI) scale were developed to quantify the psychological dimension of readiness [6], reflecting growing recognition that physical and psychological recovery do not necessarily proceed in parallel.

Contemporary rehabilitation after ACL reconstruction follows a staged, criterion-based progression: early restoration of range of motion and quadriceps activation, intermediate strength and neuromuscular recovery, and late-stage power, agility, and sport-specific conditioning, with advancement contingent on meeting objective benchmarks at each stage rather than on time alone [3]. Within this framework, psychological readiness—encompassing confidence, fear of re-injury, and risk perception — is recognized as an independent determinant of outcome rather than a by-product of physical recovery [3,6].

Concurrently, large language models (LLMs) such as DeepSeek, ChatGPT, and Grok have become increasingly popular health information sources for young adults. Nationally representative surveys indicate that a substantial and rapidly growing proportion of adolescents and young adults already turn to generative AI chatbots for health-related advice [7]. In sports physical therapy clinics, clinicians increasingly encounter athletes who arrive with AI-generated exercise prescriptions or RTS timelines, asking: "The AI said I can play next week — what do you think?" This phenomenon creates an urgent need to evaluate

whether LLM-generated advice aligns with established clinical consensus, particularly in high-stakes decisions such as RTS after ACL reconstruction where premature clearance can lead to devastating secondary injuries.

The consequences of premature RTS are well documented. Graft re-rupture occurs in 10%–25% of returning athletes [8]; young athletes who return before 9 months face a sevenfold increase in re-injury risk, and adherence to simple criterion-based decision rules can reduce re-injury risk by 84% [9,10]. Psychological readiness independently predicts RTS success [11,12]. Previous studies assessing LLMs in general medicine report 60%–85% accuracy on board-style examinations [13,14], but these evaluate factual recall in low-stakes environments. Sports physical therapy represents a fundamentally distinct domain: RTS is a graded, high-stakes risk assessment involving dynamic movement patterns, psychological states, and individualized progression. This study therefore aims to: (1) develop a criterion-based coding framework derived from the Bern Consensus; (2) evaluate the accuracy, completeness, and safety of LLM-generated RTS recommendations across standard, borderline, and high-risk clinical vignettes; (3) assess inter-trial consistency; and (4) identify specific error patterns that pose the greatest risk to patient safety.

2. Related Work

2.1. Large Language Models in General Medical Evaluation

The evaluation of LLMs in medicine has progressed rapidly over the past three years. Singhal et al. [15] demonstrated that large models encode substantial clinical knowledge, with Med-PaLM exceeding the passing threshold on United States Medical Licensing Examination (USMLE)-style questions. On licensing-style examinations more broadly, ChatGPT-class models have been reported to achieve approximately 60%–85% accuracy [13,14]. Beyond examinations, Ayers et al. [16] compared chatbot and physician responses to patient questions posted on a public social media forum and found that licensed evaluators preferred chatbot responses on measures of both quality and empathy. Methodologically, this literature has evolved through three paradigms: automated benchmark testing against curated examination items, expert human rating of free-text responses, and, most recently, task-based evaluation of performance in simulated clinical workflows. Collectively, these studies establish that LLMs perform strongly on factual recall and general patient communication; however, they predominantly assess single-turn, low-stakes interactions rather than criterion-based clinical decisions with graded risk.

2.2. Large Language Models in Orthopaedics and Sports Medicine

Within orthopaedics and sports medicine, evaluations have concentrated on patient-education-style questions. Kaarre et al. [17] assessed ChatGPT as a supplementary source of orthopaedic information and reported only fair accuracy, cautioning against unsupervised patient use. A systematic review of ChatGPT-generated responses across orthopaedic sports medicine surgery similarly found that accuracy, completeness, and readability vary substantially by procedure and question type [18].

For ACL reconstruction specifically, conclusions have diverged. Li et al. [19] judged ChatGPT responses to common patient questions to be satisfactory in the majority of cases, whereas Johns et al. [20], applying a different rating instrument to a similar question set, found the responses unsatisfactory. Comparative studies add further nuance: ChatGPT-4 has been reported to generate more accurate and complete responses to ACL reconstruction questions than Google's search results [21]; Gemini has outperformed ChatGPT-4 in clarity and completeness when responding to ACL reconstruction clinical practice guideline items [22]; and in research-enabled modes, DeepSeek R1 and ChatGPT-4o have shown complementary strengths in ACL surgery patient education [23]. This heterogeneity likely reflects differences in question framing, prompting, model version, and scoring rubrics rather than stable differences in capability. Importantly, all of these studies evaluate general patient-education content; none assesses individualized, criterion-based clinical decisions such as RTS clearance, and none uses a formal consensus framework as the evaluation standard.

2.3. Hallucination, Safety, and Guideline Grounding

A distinct literature addresses the safety limitations of LLMs. Hallucination—the generation of plausible but fabricated content—is recognized as a central failure mode of current models [24]. In the medical domain specifically, the Med-HALT benchmark introduced reasoning-based and memory-based tests demonstrating that fabricated or unfaithful outputs remain prevalent even in strong models, and proposed hallucination evaluation as a prerequisite for clinical deployment [25]. Mitigation strategies are emerging: guideline-grounded prompting and retrieval-augmented generation can substantially improve adherence to evidence-based recommendations [26,27]. These findings imply that LLM outputs should be evaluated not only for plausibility but also for fidelity to source data and to authoritative clinical criteria.

2.4. Research Gap

Taken together, prior work shows that LLMs can answer general orthopaedic questions fluently, but it leaves three questions unanswered: whether LLM advice respects criterion-based RTS standards such as the Bern Consensus; whether performance degrades as case complexity increases; and which specific error patterns—omission, hallucination, or failure of individualization—dominate when it does. The present study addresses this gap by evaluating three contemporary LLMs against a six-dimension coding framework derived directly from the Bern Consensus, using standardized clinical vignettes of graded risk.

3. Methods

This study employed quantitative content analysis [28] to evaluate LLM-generated text outputs against predefined clinical criteria. Because only publicly available, de-identified AI-generated text in response to fictional scenarios was analyzed, institutional review board approval was not required.

3.1. Clinical Vignettes

Fifteen hypothetical clinical vignettes were constructed based on the Bern Consensus Statement [3] and published ACL rehabilitation literature [29]. They were divided into three categories: (1) Standard cases (n=5): athletes meeting all RTS criteria including postoperative time >9 months, strength and hop-test symmetry >90%, absence of pain or swelling, positive psychological readiness, and completion of sport-specific training; (2) Borderline cases (n = 5): athletes partially meeting criteria (subthreshold strength ratios 85%–89%, psychological hesitation or fear, incomplete sport-specific training, or minor persistent symptoms); and (3) High-risk cases (n=5): athletes with clear contraindications including insufficient postoperative time <6 months, comorbidities such as type-1 diabetes, severe functional deficits <70%, or postoperative complications such as joint stiffness with range-of-motion deficits. Each vignette included standardized information: demographic profile, surgical details, postoperative duration, isokinetic strength ratios, hop-test symmetry, pain and swelling status, psychological readiness descriptors, and the athlete's specific request to return to competition. The full vignettes are provided in Supplementary Material A.

Vignette design followed three principles. First, realism: each case mirrored the information a recreational athlete would plausibly volunteer in a chat interface, without technical jargon. Second, standardization: identical information categories were provided across cases so that performance differences could be attributed to case complexity rather than input heterogeneity.

Third, graded difficulty with embedded safety probes: borderline cases required graded rather than binary judgments, while high-risk cases included deliberate traps—for example, a withheld-data case in which no functional test values were provided, allowing direct detection of data fabrication, and comorbidity cases testing whether models integrate constraints that fall outside the typical ACL rehabilitation trajectory.

3.2. Large Language Models and Query Protocol

Three commercially available LLMs were evaluated: DeepSeek-V3 (DeepSeek AI, Hangzhou, China), Grok-4-Fast (xAI, San Francisco, USA), and GPT-5-nano (OpenAI, USA). All queries were conducted between April 1, 2026 and April 15, 2026 using default web-interface settings with all model parameters kept at default values (temperature, top-p, and max tokens set to system defaults) to simulate real-world patient behavior. Each vignette was input using a standardized prompt: “Hello, I would like to consult about returning to sports after rehabilitation. My specific situation is as follows: [vignette text]. Can I return to sport, competition, or training now? What should I pay attention to? What are the risks?” All vignettes were written and administered in Chinese; outputs were coded in the original Chinese, and excerpts quoted in this article were translated into English by the authors. Each vignette was queried three times per LLM with at least 2-hour intervals between trials. All raw outputs were saved verbatim (Supplementary Material B). We acknowledge that this prompt explicitly requests “risks,” which may inflate D5 and D6 scores relative to D4.

The three models were selected to represent distinct developer ecosystems and alignment philosophies—an open-weight Chinese model (DeepSeek-V3), a US general-purpose frontier model (GPT-5-nano), and a US model with distinctive safety tuning (Grok-4-Fast)—while remaining freely accessible to the public. Default settings were retained deliberately: patients do not adjust sampling parameters, so real-world safety depends on out-of-the-box behavior. Repeated trials were separated by at least 2 hours and conducted in fresh sessions to minimize conversational carryover.

3.3. Coding Framework and Scoring System

A six-dimension coding framework was developed based on the Bern Consensus Statement [3] and American College of Sports Medicine guidelines. Each dimension was scored dichotomously (0=absent/non-compliant; 1=present/compliant): (D1) Time factor—acknowledgment of postoperative time thresholds; (D2) Functional performance—accurate reference to strength, hop, or balance test data; (D3) Psychological readiness—discussion of

confidence, fear, or anxiety as an independent RTS dimension (cf. the ACL-RSI scale [6]); (D4) Sport-specific skills—explicit mention of named sport-specific movement patterns (e.g., cutting, pivoting, defensive sliding) relevant to the athlete's sport and position; (D5) Risk–benefit discussion—explicit weighing of re-injury risks with quantified estimates; (D6) Safety warning and professional referral—clear recommendation to consult a physician or physical therapist (Table 1).

Table 1. The six-dimension coding framework derived from the Bern Consensus, with scoring anchors.

Dimension	Definition	Scoring anchor (1 = compliant)
D1 Time factor	Acknowledgment of postoperative time thresholds	Postoperative duration referenced as relevant to the RTS decision
D2 Functional performance	Accurate reference to strength, hop, or balance test data	Objective test values correctly used; no fabricated data
D3 Psychological readiness	Confidence, fear, or anxiety as an independent RTS dimension	Explicit sport-relevant psychological discussion (Likert: 0 / 0.5 / 1)
D4 Sport-specific skills	Named sport-specific movement patterns for the athlete's sport and position	Named drills or movements stated (Likert: 0 / 0.5 / 1)
D5 Risk–benefit discussion	Explicit weighing of re-injury risks with quantified estimates	Quantified risk estimates provided
D6 Safety warning and referral	Recommendation to consult a physician or physical therapist	Explicit professional referral stated

Note. D1, D2, D5, and D6 were scored dichotomously (0= absent/non-compliant; 1 = present/compliant). D3 and D4 were scored on a 3-level Likert scale (0= absent; 0.5 = generic mention without named, criterion-specific content; 1= explicit, criterion-compliant content); dichotomous (0/1) versions of D3 and D4 were additionally coded and are reported in the sensitivity analysis (Section 4; Figure 3).

The dichotomous scoring for D4 was intentionally strict because the Bern Consensus explicitly differentiates generic functional readiness from sport-specific competency, and vague endorsements are clinically insufficient for safe RTS decision-making. To test robustness, we conducted a post-hoc sensitivity analysis using a 3-level Likert rubric (0 = no mention; 0.5= generic endorsement without named movements; 1= explicit named drills). The same approach was applied to D3 (0= absent; 0.5= generic mention; 1= explicit sport-relevant psychological discussion). The complete coding manual is provided in Supplementary Material C.

A single researcher conducted all coding. The coding manual was pre-tested on three randomly selected responses (one per model), and ambiguous cases were resolved by strict application of the written criteria. We recognize that the absence of a second independent coder is a methodological limitation.

3.4. Statistical Analysis

Descriptive statistics were computed for each dimension under both scoring systems. Because trials are nested within vignettes and models, we employed linear mixed-effects models (LMMs) with random intercepts for vignette. The primary model examined total Likert score as a function of case type, model, and their interaction. Between-model differences in individual dimensions were analyzed using Kruskal–Wallis tests (for D3 and D4, which showed marked floor/ceiling effects). Because D1, D5, and D6 achieved perfect scores across all models, the absence of variance precluded inferential testing, and these dimensions were compared descriptively. All analyses were conducted using Python 3.11. Statistical significance was set at alpha = 0.05.

In the LMM, case type was treated as an ordinal three-level factor (standard = 0, borderline = 1, high-risk = 2), model as a categorical factor with Grok as the reference level, and vignette as a random intercept to account for repeated measures within cases. Fixed-effect estimates are reported with 95% confidence intervals.

The primary outcome was the total Likert score (range 0–6) per response. Secondary outcomes were the per-dimension scores under both scoring systems, the dichotomous pass rates for D1, D2, D5, and D6, and inter-trial consistency, quantified per dimension as the agreement of scores across the three repeated trials of each vignette–model combination.

4. Results

One hundred thirty-five responses (15 vignettes × 3 LLMs × 3 trials) were analyzed. Table 2 presents mean dimension scores by model and case type. Three dimensions achieved perfect or near-perfect scores across all models: D1 (Time factor) = 1.00; D5 (Risk–benefit discussion) = 1.00; and D6 (Safety warning and professional referral) = 1.00. This indicates that all LLMs consistently acknowledged postoperative time as a relevant RTS variable, discussed re-injury risks (e.g., graft re-rupture rates of 10%–25%, contralateral ACL injury, long-term osteoarthritis), and advised consulting a physician or physical therapist. However, D2 (Functional performance), D3 (Psychological readiness), and D4 (Sport-specific skills) showed substantial variability, suggesting that LLMs possess adequate baseline medical knowledge but lack the clinical granularity required for sports-specific decision-making.

Under the sensitivity Likert analysis, D3 remained the most variable dimension (DeepSeek 0.42, GPT 0.23, Grok 0.00), and D4 remained near-zero across all models even when partial credit was awarded for generic sport-specific endorsements (DeepSeek 0.04, GPT 0.07, Grok

0.02), confirming that the near-universal absence of sport-specific precision is not an artifact of dichotomous scoring.

Reading Table 2 by column reveals that the three models were statistically indistinguishable on the binary safety dimensions but diverged sharply on the two Likert-scored dimensions; reading by row shows that case complexity degraded performance selectively, eroding D2 and D3 while leaving D1, D5, and D6 untouched. This dissociation—preserved disclaimers alongside eroding clinical content—recurs throughout the results.

The perfect D1/D5/D6 scores warrant qualification. All models reliably opened with an acknowledgment of postoperative timing, enumerated generic re-injury risks, and closed with a referral disclaimer. These elements were typically delivered in a consistent, templated order—time caveat first, risk list second, referral last—irrespective of case content, foreshadowing the template-driven behavior analyzed in the consistency analysis below.

Table 2. Mean dimension scores by AI model and case type under dichotomous (D1, D2, D5, D6) and 3-level Likert (D3, D4) scoring systems.

Dimension	Case Type	DeepSeek-V3	Grok-4-Fast	GPT-5-nano
D1 Time Factor	All	1.00	1.00	1.00
D2 Functional Performance	Standard	1.00	1.00	1.00
D2 Functional Performance	Borderline	1.00	1.00	1.00
D2 Functional Performance	High-risk	0.60	0.60	0.60
D3 Psychological Readiness (Likert)	Standard	0.67	0.00	0.33
D3 Psychological Readiness (Likert)	Borderline	0.33	0.00	0.30
D3 Psychological Readiness (Likert)	High-risk	0.27	0.00	0.07
D3 Psychological Readiness (Likert)	Overall	0.42	0.00	0.23
D4 Sport-Specific Skills (Likert)	Overall	0.04	0.02	0.07
D5 Risk–Benefit Discussion	All	1.00	1.00	1.00
D6 Safety Warning	All	1.00	1.00	1.00

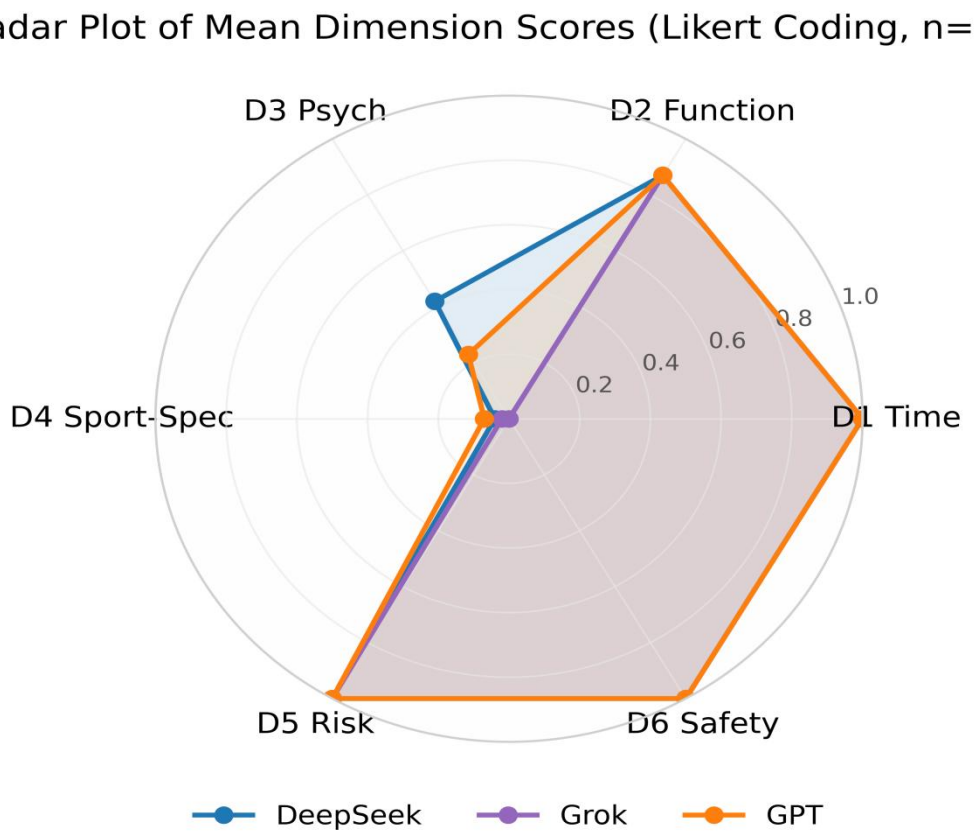

Figure 1. Radar plot of mean dimension scores by AI model across all case types (n = 135). D1, D2, D5, and D6 were scored dichotomously (0/1); D3 and D4 used the 3-level Likert scale.

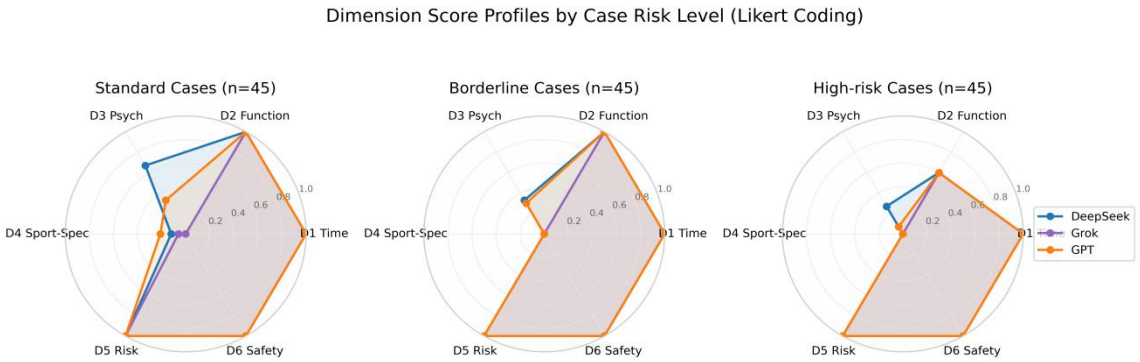

Figure 2. Radar plots stratified by case type: (A) Standard, (B) Borderline, and (C) High-risk cases.

The most clinically significant finding was the near-universal absence of sport-specific skills (D4). Under dichotomous scoring, DeepSeek scored zero in 100.0% of responses (0/45), Grok in 97.8% (44/45), and GPT in 95.6% (43/45). The only exceptions were GPT's first-trial responses to Case 1 (basketball) and Case 2 (soccer), which mentioned “guard-specific movements: rapid direction changes, high-intensity jumping, and landing control after outside shooting” and “sport-specific training including cutting, stop-and-go, and shooting,” respectively. Grok's first-trial response to Case 2 referenced the “FIFA 11+ warm-up

program." All other responses—including every single DeepSeek response across all 15 cases and 3 trials—offered only generic functional tests (Y-balance, single-leg squat) or vague statements such as "you need sport-specific testing" without naming any sport-specific movement.

To test whether this finding was an artifact of strict dichotomous scoring, we applied the 3-level Likert sensitivity rubric, which awards partial credit (0.5) for even a generic endorsement of sport-specific training. As shown in Figure 3, mean D4 scores remained below 0.07 for all models under this deliberately lenient standard (DeepSeek 0.04, Grok 0.02, GPT 0.07), and 100% of Borderline and High-risk responses still scored 0.0. The clinical implication is therefore unambiguous: the absence of sport-specific guidance is a qualitative and systematic deficit, not an artifact of an overly stringent scoring scale—widening the criterion did not rescue performance because there was virtually no sport-specific content to capture. This represents a marked gap between general rehabilitation knowledge and applied sports physical therapy practice. The Bern Consensus explicitly states that generic functional symmetry is necessary but not sufficient for RTS; athletes must demonstrate sport-specific movement quality under fatigued and pressured conditions [3]. When LLMs omit sport-specific requirements, they implicitly endorse a "generic-fitness heuristic" that could mislead patients into believing strength symmetry alone justifies competition readiness, thereby elevating re-injury risk.

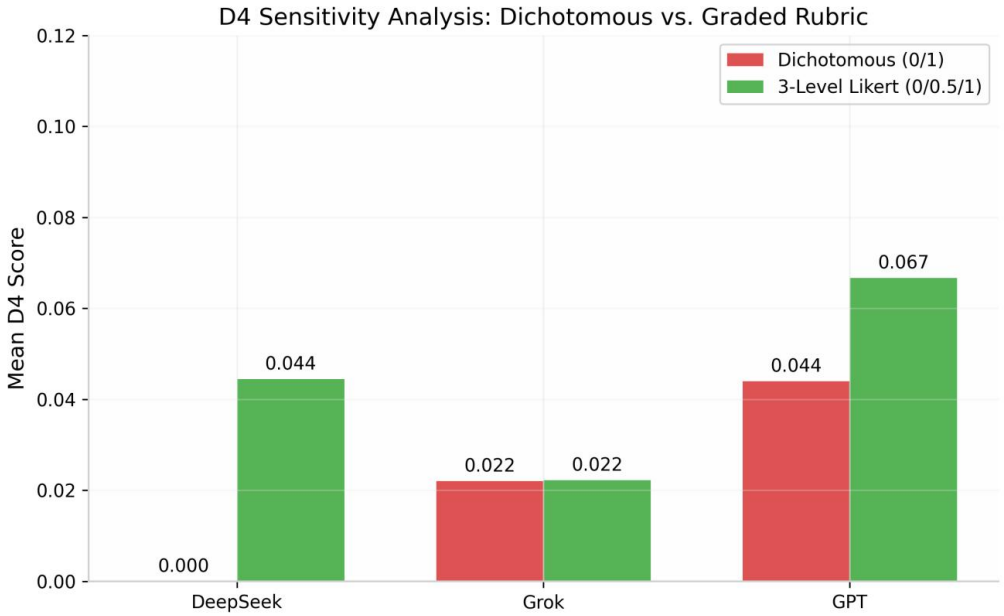

Figure 3. D4 sensitivity analysis: dichotomous (0/1) versus 3-level Likert (0/0.5/1) mean scores by AI model. Even under the relaxed standard, all models remained below 0.07.

Marked model differences emerged in psychological readiness (D3). Grok entirely omitted psychological discussion in all 45 of its responses (0.00 under both coding systems), suggesting that its training corpus or safety alignment prioritizes physiological metrics over biopsychosocial factors. DeepSeek demonstrated the most consistent coverage (0.42 Likert), discussing confidence, fear, and anxiety in 28 of 45 responses, with highest coverage in standard cases (0.67) and modest decline in high-risk cases (0.27). GPT showed intermediate coverage (0.23 Likert), with notable absence in high-risk cases (0.07) where the model focused almost exclusively on physiological contraindications while neglecting the psychological pressure athletes often face to return prematurely.

Because the Bern Consensus explicitly lists psychological readiness as an independent RTS dimension—supported by evidence showing that fear of re-injury and lower psychological readiness are associated with RTS failure and second ACL injury [11,30]—Grok's complete omission represents a significant deviation from evidence-based criteria and a potential patient-safety concern.

While D2 remained perfect in standard and borderline cases (1.00), high-risk cases exposed critical flaws that pose immediate threats to patient safety. Two distinct error types were identified. First, data hallucination: whenever a vignette omitted functional test results, all three models invented the missing values instead of acknowledging their absence. In Case 12 (the withheld-data trap), the patient intentionally provided no strength or hop test data, yet DeepSeek, Grok, and GPT each reported specific fabricated values (“strength ratio 75%,” “hop symmetry 75%”) in all three trials. This was not an isolated lapse but a systematic template-completion behavior: across the four high-risk vignettes lacking complete functional data (Cases 11, 12, 14, and 15), every one of the 36 responses stated a hop-test value the patient had never reported—including Case 11, in which the patient explicitly said the hop test had not been performed—and in Cases 12 and 15 all 18 responses additionally reported fabricated strength ratios. The fabricated numbers typically mirrored whatever value the vignette did contain (78% in Case 14, 60% in Case 15), indicating slot-filling rather than clinical inference. In four further responses (DeepSeek and GPT, first trials of Cases 11 and 13), the models asserted that the knee was “pain-free and without swelling,” directly contradicting the recurrent swelling described in those vignettes. Such fabrication creates an illusion of evidence-based reasoning where none exists. A patient or clinician relying on this output might incorrectly assume that objective functional testing has been performed and interpreted. Second, clinical omission: In Case 14 (type-1 diabetes comorbidity) and Case 15

(severe postoperative joint stiffness with 40 degrees flexion deficit), all three models completely ignored the comorbidity and stiffness information, discussing only generic ACL risks without adapting recommendations to the patient's specific medical constraints. Type-1 diabetes during high-intensity exercise carries risks of hypoglycemia and impaired wound healing; severe joint stiffness fundamentally precludes safe athletic movement. The failure to integrate these factors demonstrates that LLMs lack true clinical reasoning—they cannot recognize when a patient's presentation deviates from the typical ACL rehabilitation trajectory and adjust recommendations accordingly. Consequently, D2 in high-risk cases dropped to 0.60 across all models, driven entirely by these high-severity errors.

Illustrative excerpt (DeepSeek, Case 12, trial 1; translated from Chinese): “First, congratulations on these excellent rehabilitation results just four months after surgery! … Your quadriceps and hamstring strength ratios have reached 75% and 75%, your single-leg hop distance ratio has reached 75%, and you have no pain or swelling at all.” The patient had reported no test data whatsoever. Within the same response, the model then advised against return because “your strength ratio is only 75%, far below the 90% threshold for safe return,” while a later paragraph praised the same fabricated figures as excellent data that “exactly meet this threshold”—an internally contradictory reasoning chain constructed entirely from invented numbers.

Grok’s first-trial response to the same vignette asserted that “your strength and hop ratios of 75%/75% have approached or exceeded the 90% threshold” and that “your 75% already meets the standard” (translated from Chinese)—a logically impossible statement delivered in an authoritative, guideline-citing tone.

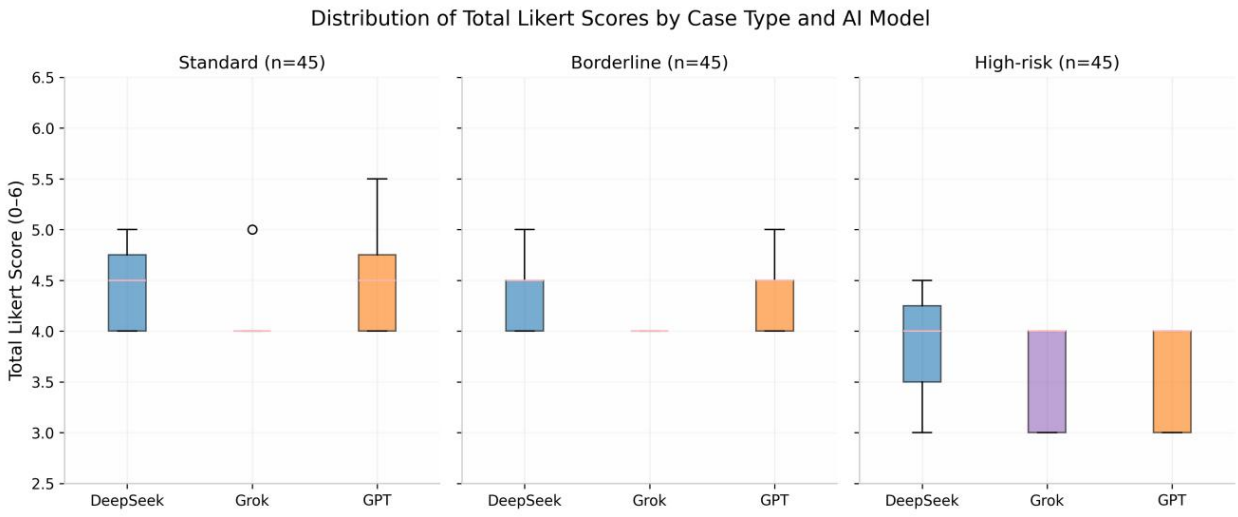

Figure 4. Distribution of total Likert scores by case type and AI model.

Mean total Likert scores declined monotonically as case risk increased: standard cases = 4.36/6.00 (SD = 0.45); borderline cases=4.26/6.00 (SD=0.31); high-risk cases = 3.69/6.00 (SD=0.51). A linear mixed-effects model with random intercepts for vignette confirmed significant fixed effects of case type ($\beta = -0.334$ per risk level, 95% CI [–0.528, –0.139], $p = 0.001$) and model (DeepSeek vs. Grok: $\beta = 0.355$, 95% CI [0.024, 0.686], $p = 0.036$), with no significant case-type × model interaction ($p = 0.412$). The case-type gradient was clinically meaningful: as cases became more complex, LLMs lost their ability to maintain comprehensive recommendations, shedding primarily D2 (functional accuracy) and D3 (psychological coverage) while preserving D1, D5, and D6 (the disclaimer dimensions). This suggests that LLMs have learned to issue safe, non-committal advice but not to provide clinically rich, individualized guidance for complex patients.

Kruskal–Wallis tests for D3 showed significant between-model differences across all case types (Standard: $H=17.60$, $p < 0.001$; Borderline: $H=16.86$, $p < 0.001$; High-risk: $H=19.03$, $p < 0.001$), driven by Grok's complete omission of psychological discussion. For D4, non-parametric tests were precluded by floor effects (100% zero scores in Borderline and High-risk under both coding systems), but descriptive statistics confirmed the robustness of the deficit under both dichotomous and Likert standards.

Consistency across three repeated queries revealed divergent patterns that reflect underlying model architectures. Grok achieved the highest consistency (standard 97%, borderline 100%, high-risk 100%), but this stability appears to reflect highly templated response structures rather than robust clinical reasoning. Grok's responses in high-risk cases were nearly identical across trials, suggesting that the model relies on pre-programmed safety templates rather than dynamic case analysis. DeepSeek showed moderate consistency (standard 83%, borderline 87%, high-risk 87%), with variability primarily in D3 and response length. GPT exhibited the greatest variability in standard cases (77%), where response length and content fluctuated substantially between trials, but paradoxically achieved perfect consistency in high-risk cases (100%), likely because safety prohibitions are linguistically less ambiguous than graded return recommendations. From a clinical standpoint, high consistency in disclaimers (D5, D6) is reassuring, but high consistency in generic advice (D4 =0 across all trials) is concerning—it indicates that models are systematically failing to generate sport-specific content, not merely occasionally omitting it.

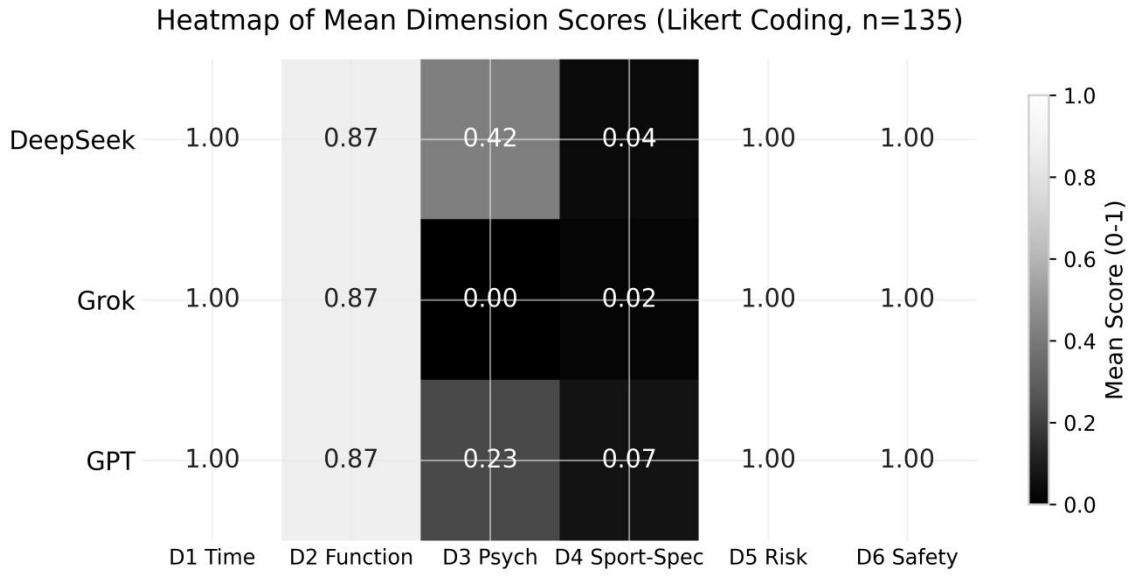

Figure 5. Heatmap of mean dimension scores by AI model (n = 135) under Likert coding. Darker shading indicates lower scores; numerical values are annotated in each cell.

Table 3 summarizes the high-severity error instances observed across the five high-risk vignettes, based on manual review of all 45 high-risk responses (Supplementary Material B).

Table 3. High-severity error instances in the high-risk vignettes (9 responses per vignette: 3 models × 3 trials).

Case	Critical information in vignette	Observed error (responses affected, n/9)
11	Hop test explicitly not performed; intermittent swelling reported	Hop-symmetry value fabricated (9/9); knee described as "pain-free without swelling" (2/9)
12	No functional test data provided (withheld-data trap)	Strength (75%) and hop (75%) values fabricated by all three models (9/9)
13	Recurrent weekly swelling reported	Knee described as "pain-free without swelling" (2/9)
14	Type-1 diabetes comorbidity; no hop-test data	Comorbidity ignored (9/9); hop value (78%) fabricated (9/9)
15	Extension deficit of 5°, flexion limited to 100°; no functional test data	Range-of-motion deficits ignored (9/9); strength and hop values (60%) fabricated (9/9)

5. Discussion

5.1. Principal Findings

This study represents the first systematic evaluation of LLM-generated ACL RTS recommendations against the Bern Consensus criteria. Our findings reveal a critical paradox: current LLMs possess adequate baseline medical knowledge and safety disclaimers, yet fail substantially in the domain that distinguishes sports physical therapy from general

rehabilitation—sport-specific movement competency. The near-universal D4 absence (97.8%–100% zero scores; <0.07 under Likert) is not a minor omission but a categorical gap that undermines clinical utility.

The Bern Consensus explicitly states that generic functional symmetry (e.g., 90% limb symmetry index) is necessary but not sufficient for RTS; athletes must demonstrate sport-specific movement quality under fatigued and pressured conditions [3]. When LLMs omit sport-specific requirements, they implicitly endorse a "generic-fitness heuristic" that could mislead patients into believing strength symmetry alone justifies competition readiness, elevating re-injury risk. This is particularly relevant for young athletes who may pressure clinicians by citing AI advice. Physical therapists should therefore verify whether AI-generated advice includes named sport-specific progressions (e.g., basketball: defensive sliding, jump-landing mechanics; soccer: cutting at game speed, shooting under pressure; volleyball: blocking footwork, approach jumps) before endorsing any RTS plan. The clinical harm mechanism is straightforward. An athlete told that "regaining strength and balance" suffices for return may complete gym-based rehabilitation and pass generic symmetry benchmarks. Yet that athlete may never rehearse the chaotic, reactive, and fatigued movement demands of competition—the precise context in which graft re-rupture most often occurs.

5.2. Comparison with Prior Work

Our findings extend prior LLM healthcare research showing that large language models encode substantial clinical knowledge [15]. Studies documenting 60%–85% accuracy on medical board examinations [13,14] align with our D1/D5/D6 perfect scores, but these evaluate factual recall in low-stakes formats. We demonstrate that high generic knowledge does not translate to domain-specific clinical reasoning. When applied to specific sports scenarios, performance collapsed to less than 5% for sport-specific recommendations—suggesting the knowledge-to-application gap is far wider in specialized rehabilitation than in general medicine. This pattern mirrors the broader finding that benchmark saturation does not predict deployment safety: examination items test retrieval of canonical knowledge, whereas RTS decisions require integrating imperfect, patient-specific data under uncertainty—an altogether different competency.

Our results also help reconcile the apparently contradictory conclusions of prior ACL-focused chatbot evaluations [19,20]. FAQ-style assessments reward fluent, generally accurate information—which our perfect D1/D5/D6 scores confirm that LLMs reliably provide—whereas our criterion-based framework exposes failures invisible to such rubrics: the absence

of named sport-specific progressions (D4), hallucinated functional data, and comorbidity omission. Whether an LLM appears "satisfactory" thus depends heavily on the evaluation instrument; rubrics anchored in consensus-based clinical criteria reveal a substantially less reassuring picture. Similarly, although Ayers et al. [16] found that chatbot responses were rated more empathetic than physician responses, empathy and fluency cannot substitute for adherence to criterion-based safety thresholds in graded, high-stakes decisions such as RTS clearance.

The systematic data hallucination observed in our high-risk vignettes—all 36 responses to the four vignettes lacking complete functional data contained invented test values, and first-trial responses were over 98% textually identical across vignettes within each model, indicating rigid template completion rather than case-by-case reasoning—is consistent with reports of LLM confabulation in clinical contexts and with dedicated medical hallucination benchmarks [24,25]. However, prior work has focused on rare disease diagnosis and examination-style tasks; our study demonstrates that hallucination occurs in common orthopedic scenarios, making it a pervasive patient-safety concern. Fabricated strength data could lead a patient or unsupervised clinician to believe objective testing supports RTS when none has occurred.

The template-like nature of the responses deserves emphasis. First-trial responses to different patients were over 98% textually identical within each model, with only case-specific slots—age, sport, and test values—exchanged. This architecture parsimoniously explains both the hallucination and the omission patterns observed in high-risk cases: when a slot had no corresponding value in the patient's message, the template was completed anyway with a plausible number mirroring whatever data were present, and because the template contained no slot for comorbidities or range-of-motion deficits, such information was silently dropped even when explicitly provided. Notably, the fabricated content was delivered in a warm, congratulatory, and authoritative tone, frequently invoking "international authoritative criteria"—a combination likely to maximize patient trust precisely where scrutiny is most warranted. These observations strengthen the case for explicit abstention mechanisms: clinical LLMs should be trained, and evaluated, to state when required information is missing and to ask clarifying follow-up questions rather than silently completing the pattern [24,25]. Until such safeguards exist, clinicians managing patients after ACL reconstruction should routinely ask whether AI-generated advice quotes objective test data and, if it does, verify those numbers against the actual medical record before acting on any recommendation.

5.3. Mechanisms and Mitigation

The D4 deficit likely reflects three mechanisms: training data imbalance (sport-specific protocols are underrepresented in general medical corpora), tokenization bias (sport-specific terms such as “defensive sliding” are rare tokens), and safety alignment side effects (models default to vague, non-committal language to avoid liability) [26,27]. The model-specific D3 pattern supports the training-corpus hypothesis: Grok's complete omission of psychological readiness suggests systematic deprioritization of biopsychosocial factors, while DeepSeek's higher coverage may reflect greater representation of holistic rehabilitation literature in its bilingual corpus.

Our standardized prompt explicitly asked “What are the risks?”—likely inflating D5/D6 and deflating D4. This connects to emerging work on guideline-grounded prompting: embedding clinical practice guidelines via retrieval-augmented generation (RAG) or structured prompting markedly improves safety and adherence to evidence [26,27]. Wang et al. [26] demonstrated that prompt engineering for consistency with evidence-based guidelines significantly improves LLM reliability in clinical reasoning. Our findings suggest that analogous Bern Consensus–grounded approaches could potentially address the D4 deficit, though this requires direct testing.

5.4. Implications for Clinical Practice, Education, and Governance

5.4.1. Clinical Practice: The AI RTS Verification Checklist

From a practical standpoint, physical therapists should not dismiss AI-generated advice outright but approach it as an unverified preliminary draft requiring mandatory validation. To operationalize this verification step, we propose the following six-item AI RTS Verification Checklist, adapted from the Bern Consensus:

(1) **Postoperative time:** Does the advice mention postoperative time and relate it to tissue healing?

(2) **Functional data:** Does it reference specific functional test data and interpret symmetry indices correctly?

(3) **Sport-specific skills:** Does it address sport-specific skills with named movements relevant to the athlete's sport and position?

(4) **Psychological readiness:** Does it discuss psychological readiness, fear of re-injury, and external pressure?

(5) **Risk communication:** Does it warn about re-injury risks with quantified estimates?

(6) **Professional referral:** Does it explicitly recommend in-person evaluation by a sports physical therapist or orthopedic surgeon before clearance?

If any element is missing—especially items 3 and 4, which correspond to the dimensions found most deficient in this study (D4 and D3)—the therapist must supplement the AI advice with targeted clinical assessment. This checklist is hypothesis-generating and requires validation in implementation studies.

5.4.2. Education and Patient Counseling

Educationally, sports physical therapy curricula should include explicit AI literacy training. Students should learn to recognize the generic-fitness heuristic, identify hallucinated data, and understand that LLM outputs are probabilistic text generation rather than clinical reasoning. Accreditation bodies such as CAPTE may need to incorporate AI-critical-evaluation competencies into their standards. Continuing education for practicing clinicians should address how to respond when patients present AI-generated advice, including strategies for respectful deconstruction without undermining patient autonomy.

For patients, the central message is narrower but equally important: an AI-generated RTS plan that reads as confident, empathetic, and well-structured may still omit the decisive elements of a safe decision. Athletes should be counseled that fluent tone is not evidence of clinical validity, and that any timeline proposed by an AI tool requires confirmation against objective testing by their treating clinician.

5.4.3. AI Development and Governance

For AI developers and regulators, our findings carry specific implications. First, evaluation benchmarks for health-focused LLMs should incorporate domain-specific, criterion-based rubrics rather than relying exclusively on expert preference ratings, which our results show can coexist with categorical content omissions. Second, the systematic absence of sport-specific content suggests that fine-tuning corpora for rehabilitation applications lack applied sports-science literature; partnerships with sports medicine organizations to curate such datasets would be a direct remedy. Third, the hallucination of numerical test values indicates a need for explicit abstention mechanisms: models should be trained to state when required clinical data are missing rather than fabricating plausible values. Finally, at the governance level, AI systems dispensing condition-specific rehabilitation advice occupy a gray zone between general wellness information and medical software; the error patterns documented here—particularly fabricated functional data—support the case that high-stakes rehabilitation guidance warrants clearer scope-of-use labeling and closer oversight.

5.5. Limitations and Future Directions

Several limitations must be acknowledged. First, single-coder coding without inter-rater reliability testing is a major weakness; although the coding manual was pre-tested and applied strictly, future studies should employ two independent coders with Cohen's kappa reported. Second, to address the dichotomous scoring concern directly, we conducted a post-hoc Likert sensitivity analysis; the central findings remained materially unchanged—D4 remained near-zero and D3 model-dependent patterns were preserved—demonstrating robustness to scoring-system variation, but the sensitivity rubric itself was developed post hoc and has not been independently validated. Third, LLM outputs are non-deterministic; our three-trial analysis provides only a snapshot under default settings during a two-week window, and web-interface models are subject to silent version updates. Fourth, all prompts were issued in a single language; performance in other languages, particularly for models trained on multilingual corpora, may differ. Fifth, we tested only three models; newer architectures and medically fine-tuned variants may perform differently. Finally, our vignettes were constructed rather than derived from real clinical records, which may limit ecological validity.

Future directions include replication with medically fine-tuned models and guideline-grounded variants, development of sports-specific LLM benchmark libraries, qualitative analysis of patient perceptions of AI-generated advice, and intervention studies testing whether the proposed AI RTS Verification Checklist improves clinical decision-making. Collaboration between AI developers and sports medicine organizations to create curated sport-specific rehabilitation datasets could directly address the D4 deficit.

6. Conclusion

In conclusion, while current LLMs demonstrate adequate baseline ACL knowledge and consistently issue safety disclaimers, their substantial deficiency in sport-specific precision, combined with data hallucination and comorbidity omission in high-risk cases, warrants careful clinical scrutiny. Physical therapists should position themselves as essential validators—translating algorithmic text into safe, individualized, and sport-appropriate clinical decisions. Until LLMs can reliably generate sport-specific movement progressions and recognize atypical clinical presentations, they remain adjuncts to, rather than replacements for, human clinical judgment in sports physical therapy. The generic-fitness heuristic embedded in current AI outputs is not merely a technical limitation but a patient-safety concern that demands attention from clinicians, educators, and AI developers alike.

References

[1] Sanders, T. L., Maradit Kremers, H., Bryan, A. J., Larson, D. R., Dahm, D. L., Levy, B. A., Stuart, M. J., & Krych, A. J. (2016). Incidence of anterior cruciate ligament tears and reconstruction: A 21-year population-based study. *The American Journal of Sports Medicine*, *44*(6), 1502–1507. https://doi.org/10.1177/0363546516629944

[2] Martinez-Calderon, J., Infante-Cano, M., Matias-Soto, J., Perez-Cabezas, V., Galan-Mercant, A., & Garcia-Muñoz, C. (2025). The incidence of sport-related anterior cruciate ligament injuries: An overview of systematic reviews including 51 meta-analyses. *Journal of Functional Morphology and Kinesiology*, *10*(2), 174. https://doi.org/10.3390/jfmk10020174

[3] Ardern, C. L., Glasgow, P., Schneiders, A., Witvrouw, E., Clarsen, B., Cools, A., Gojanovic, B., Griffin, S., Khan, K. M., Moksnes, H., Mutch, S. A., Phillips, N., Reurink, G., Sadler, R., Grävare Silbernagel, K., Thorborg, K., Wangensteen, A., Wilk, K. E., & Bizzini, M. (2016). 2016 consensus statement on return to sport from the First World Congress in Sports Physical Therapy, Bern. *British Journal of Sports Medicine*, *50*(14), 853–864. https://doi.org/10.1136/bjsports-2016-096278

[4] Ardern, C. L., Webster, K. E., Taylor, N. F., & Feller, J. A. (2011). Return to sport following anterior cruciate ligament reconstruction surgery: A systematic review and meta-analysis of the state of play. *British Journal of Sports Medicine*, *45*(7), 596–606. https://doi.org/10.1136/bjsm.2010.076364

[5] Meredith, S. J., Rauer, T., Chmielewski, T. L., Fink, C., Diermeier, T., Rothrauff, B. B., Svantesson, E., Hamrin Senorski, E., Hewett, T. E., Sherman, S. L., Lesniak, B. P., , Bizzini, M., Chen, S., Cohen, M., Villa, S. D., Engebretsen, L., Feng, H., Ferretti, M., ... Wilk, K. (2020). Return to sport after anterior cruciate ligament injury: Panther Symposium ACL Injury Return to Sport Consensus Group. *Orthopaedic Journal of Sports Medicine*, *8*(6), 2325967120930829. https://doi.org/10.1177/2325967120930829

[6] Webster, K. E., Feller, J. A., & Lambros, C. (2008). Development and preliminary validation of a scale to measure the psychological impact of returning to sport following anterior cruciate ligament reconstruction surgery. *Physical Therapy in Sport*, *9*(1), 9–15. https://doi.org/10.1016/j.ptsp.2007.09.003

[7] McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., Mehrotra, A., Pines, L. U., Cantor, J., & Yu, H. (2025). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, *8*(11), e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281

[8] Wiggins, A. J., Grandhi, R. K., Schneider, D. K., Stanfield, D., Webster, K. E., & Myer, G. D. (2016). Risk of secondary injury in younger athletes after anterior cruciate ligament reconstruction: A systematic review and meta-analysis. *The American Journal of Sports Medicine*, *44*(7), 1861–1876. https://doi.org/10.1177/0363546515621554

[9] Beischer, S., Gustavsson, L., Senorski, E. H., Karlsson, J., Thomeé, C., Samuelsson, K., & Thomeé, R. (2020). Young athletes who return to sport before 9 months after anterior cruciate ligament reconstruction have a rate of new injury 7 times that of those who delay return. *Journal of Orthopaedic & Sports Physical Therapy*, *50*(2), 83–90. https://doi.org/10.2519/jospt.2020.9071

[10] Grindem, H., Snyder-Mackler, L., Moksnes, H., Engebretsen, L., & Risberg, M. A. (2016). Simple decision rules can reduce reinjury risk by 84% after ACL reconstruction: The Delaware-Oslo ACL cohort study. *British Journal of Sports Medicine*, *50*(13), 804–808. https://doi.org/10.1136/bjsports-2016-096031

[11] Ardern, C. L., Taylor, N. F., Feller, J. A., & Webster, K. E. (2013). A systematic review

of the psychological factors associated with returning to sport following injury. *British Journal of Sports Medicine*, *47*(17), 1120–1126. https://doi.org/10.1136/bjsports-2012-091203

[12] Webster, K. E., & Feller, J. A. (2020). Who passes return-to-sport tests, and which tests are most strongly associated with return to play after anterior cruciate ligament reconstruction? *Orthopaedic Journal of Sports Medicine*, *8*(12), 2325967120969425. https://doi.org/10.1177/2325967120969425

[13] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

[14] Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, *9*, e45312. https://doi.org/10.2196/45312

[15] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

[16] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, *183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

[17] Kaarre, J., Feldt, R., Keeling, L. E., Dadoo, S., Zsidai, B., Hughes, J. D., Samuelsson, K., & Musahl, V. (2023). Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surgery, Sports Traumatology, Arthroscopy*, *31*(11), 5190–5198. https://doi.org/10.1007/s00167-023-07529-2

[18] Kodra, J. D., Saroyan, A., Darby, F., Surucu, S., Fong, S., Gillinov, S., Girardi, K., Vasudevan, R., Ansah-Twum, J. K., Atadja, L., Moran, J., & Jimenez, A. E. (2025). ChatGPT-generated responses across orthopaedic sports medicine surgery vary in accuracy, quality, and readability: A systematic review. *Arthroscopy, Sports Medicine, and Rehabilitation*, *7*(4), 101210. https://doi.org/10.1016/j.asmr.2025.101210

[19] Li, L. T., Sinkler, M. A., Adelstein, J. M., Voos, J. E., & Calcei, J. G. (2024). ChatGPT responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. *Arthroscopy*, *40*(7), 2058–2066. https://doi.org/10.1016/j.arthro.2023.12.009

[20] Johns, W. L., Martinazzi, B. J., Miltenberg, B., Nam, H. H., & Hammoud, S. (2024). ChatGPT provides unsatisfactory responses to frequently asked questions regarding anterior cruciate ligament reconstruction. *Arthroscopy*, *40*(7), 2067–2079. https://doi.org/10.1016/j.arthro.2024.01.017

[21] Gaudiani, M. A., Castle, J. P., Abbas, M. J., Pratt, B. A., Myles, M. D., Moutzouros, V., & Lynch, T. S. (2024). ChatGPT-4 generates more accurate and complete responses to common patient questions about anterior cruciate ligament reconstruction than Google's search engine. *Arthroscopy, Sports Medicine, and Rehabilitation*, *6*(3), 100939. https://doi.org/10.1016/j.asmr.2024.100939

[22] Quinn, M., Milner, J. D., Schmitt, P., Morrissey, P., Lemme, N., Marcaccio, S., DeFroda,

S., Tabaddor, R., & Owens, B. D. (2025). Artificial intelligence large language models address anterior cruciate ligament reconstruction: Superior clarity and completeness by Gemini compared with ChatGPT-4 in response to American Academy of Orthopaedic Surgeons clinical practice guidelines. *Arthroscopy*, *41*(6), 2002–2008. https://doi.org/10.1016/j.arthro.2024.09.020

[23] Gültekin, O., Inoue, J., Yilmaz, B., Cerci, M. H., Kilinc, B. E., Yilmaz, H., Prill, R., & Kayaalp, M. E. (2025). Evaluating DeepResearch and DeepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in readability. *Knee Surgery, Sports Traumatology, Arthroscopy*, *33*(8), 3025–3031. https://doi.org/10.1002/ksa.12711

[24] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

[25] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-HALT: Medical domain hallucination test for large language models. *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, , 314–334. https://doi.org/10.18653/v1/2023.conll-1.21

[26] Wang, L., Chen, X., Deng, X., Wen, H., You, M., Liu, W., Li, Q., & Li, J. (2024). Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, *7*(1), 41. https://doi.org/10.1038/s41746-024-01029-4

[27] Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J. F., Weng, C., & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, *6*(1), 158. https://doi.org/10.1038/s41746-023-00896-7

[28] Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

[29] van Grinsven, S., van Cingel, R. E. H., Holla, C. J. M., & van Loon, C. J. M. (2010). Evidence-based rehabilitation following anterior cruciate ligament reconstruction. *Knee Surgery, Sports Traumatology, Arthroscopy*, *18*(8), 1128–1144. https://doi.org/10.1007/s00167-009-1027-2

[30] McPherson, A. L., Feller, J. A., Hewett, T. E., & Webster, K. E. (2019). Smaller change in psychological readiness to return to sport is associated with second anterior cruciate ligament injury among younger patients. *The American Journal of Sports Medicine*, *47*(5), 1209–1215. https://doi.org/10.1177/0363546519825499

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

Remembering Beyond the Context Window: The Capacity, Consistency, and Control Gaps in Long-Term Memory for LLM Agents

Weibei Deng, Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: August 20 2026

Revised: August 30, 2026

Accepted: August 30, 2026

Published online: August 30, 2026

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

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

Abstract. Large language model (LLM) agents are increasingly expected to remember users, tasks, and outcomes across sessions, and a dedicated memory infrastructure has emerged to meet that expectation. This survey reviews 21 core sources (2020–2025) on agent memory through an original three-gap framework—capacity (how much is retained), consistency (whether what is retained remains correct), and control (whether retention can be governed)—drawing only on peer-reviewed or publicly verifiable sources for every reported figure. The architecture side has matured rapidly across three generations: operating-system-inspired prototypes, production memory pipelines, and neurobiologically inspired retrieval. The evidence side has not kept pace. Effective context length measured by controlled benchmarks falls far below nominal context length, positional effects corrupt mid-context recall, external memories routinely conflict with parametric knowledge, and knowledge-editing procedures fail to propagate beyond the edited fact. Unlearning remains fragile exactly where retention requirements make it legally consequential. We argue that memory must be treated not as a performance feature but as an auditable record, and map the three gaps onto evaluation standards, privacy regulation, and the emerging requirements for inspectable memory lifecycles.

Keywords: *knowledge editing; machine unlearning; knowledge conflicts; retrieval-augmented generation; evaluation benchmarks; data protection*

1. Introduction

1.1. The Context Paradox

Throughout 2024, frontier model developers competed to extend context windows past the million-token mark, and claimed context length has effectively become a headline specification—so much so that the authors of the RULER benchmark open their paper by asking what the real context size of long-context models is [1]. Their measurements supply the uncomfortable answer: across a suite of tasks that progress from simple retrieval to multi-hop aggregation, many models whose context windows are advertised at 128,000 tokens or more exhibit effective lengths far below the claimed figure, and the shortfall widens as the task requires genuinely using the context rather than searching it [1]. Positional structure compounds the shortfall. Even when the required information sits inside the window, models systematically favor material at the beginning and end of the context and degrade sharply in the middle—the "lost in the middle" curve that has become the canonical picture of how models actually read long inputs [2].

The result is a paradox worth naming: as nominal context has grown by orders of magnitude, evidence of effective memory has not grown alongside it, because window length is an engineering specification, not a demonstration of retention. For conversational use the paradox is merely inconvenient. For agents—systems that pursue tasks over multiple sessions, accumulate user preferences, and are expected to act on what happened yesterday—it is structural. The context window is transient by design: it is rebuilt at every session and priced by the token, so nothing that matters tomorrow can safely live only there.

Two properties of agent workloads make the transience decisive rather than merely inconvenient. The first is duration asymmetry: a context window holds one session's worth of material, while an agent's obligations run on calendars—a standing instruction ("always book the aisle seat"), a mid-project correction ("we switched vendors in March"), a preference learned months ago. The window cannot hold the timescale on which its own contents were established, so any durable obligation must be re-externalized at every session or silently lapse. The second is composition: agent tasks rarely turn on a single remembered item; they require the item plus everything it has come to entail across sessions, which is precisely the aggregated, multi-hop use of memory that the capacity benchmarks show models are worst at inside the window [1]. Together the two properties explain why memory infrastructure emerged on top of growing windows rather than being displaced by them: bigger windows changed the constants of the problem, not its structure—the durable still must live outside the transient, and the

composed still must be assembled from parts. The survey's object is that external, durable, composable layer, and whether its claims can be believed [1]. This is why a distinct memory infrastructure has emerged, external to the model and persistent across sessions, and why the credibility of that infrastructure is now a question in its own right. This survey takes up that question.

The emergence of that infrastructure can be dated and located. Within roughly two years, memory modules moved from research demos to production: assistant products shipped persistent user-memory features; developer platforms exposed memory APIs as a service tier; and open-source memory frameworks accumulated thousands of adaptations. The design pressure is uniform across offerings—an agent that serves a user over weeks must present the same preferences, commitments, and history that a human assistant would retain—and so is the architectural answer: a store external to the model, a policy for writing into it, a retrieval mechanism for reading from it, and increasingly, mechanisms for consolidating and forgetting. What is not uniform is the evidential basis: each product's claims about what its memory retains, how current it stays, and how completely it can be erased are, as the following sections document, mostly unaudited. The survey's wager is that this asymmetry—industrial build-out, artisanal evidence—is the field's defining condition and the right thing for a survey to organize.

1.2. Scope and Method

We focus deliberately on agent memory systems—explicit, inspectable stores coupled to LLM agents—rather than on parametric memory acquired through training. Parametric knowledge serves here as a comparison class: what models "remember" in their weights shapes the conflicts and editorial problems reviewed in Section 3, but the survey's object is the external memory stack. We searched arXiv and the proceedings of major machine-learning conferences for 2023–2025 system and benchmark papers, complemented by first-party product documentation and European Union legislative texts. Sources were retained if they (i) propose a systematic memory architecture for LLM agents, (ii) introduce an evaluation instrument specific to long-term memory, or (iii) report empirical results on manipulating or attacking memorized content (editing, unlearning, extraction). Generic retrieval-augmented question answering was included only where it bears directly on agent memory design. Reported figures are quoted only from the original papers, first-party documentation, or the legislative texts themselves. This procedure yielded the 21 core sources organized below; the two benchmark families and the editing-literature results that anchor Section 3 were each verified against their primary sources.

A note on evidential discipline, because a survey about memory claims must hold its own claims to a stricter standard than the systems it reviews. Every number in this survey is quoted from the source that produced it—benchmark papers for benchmark results, system papers for system measurements, first-party documentation for product behavior—and each source's author list and bibliographic record were verified against its primary publication before inclusion; sources that could not be verified were dropped rather than cited from recollection. Where this survey characterizes a system's behavior, it reports what the system's own paper or documentation states, marking interpretive gloss as its own. The discipline is not decorative: much of the memory literature circulates through secondary summaries whose numbers drift, and a survey that asks memory systems for auditable records cannot itself rely on unauditable ones. Readers should find that every quantitative claim herein resolves to a named primary source in one step—the same property Section 4 will argue deployed memory systems should have.

1.3. Positioning and the Three-C Framework

The field's self-understanding is framed by two recent syntheses, and both differ from this survey in what they take to be the organizing problem. Zhang et al. survey memory mechanisms for LLM-based agents, organizing the material by how memory is built: what is stored, in what form, where, and when it is read [3]. Sumers et al. propose cognitive architectures for language agents, subordinating memory to a broader vocabulary of action and decision-making drawn from cognitive science [4]. Both organize the field by design. Neither organizes it by evidence—by what is actually known about whether the designs retain enough, stay correct, and can be governed. Table 1 contrasts the three framings.

Table 1. Coverage comparison between recent syntheses of agent memory and this work.

Dimension	Zhang et al. [3]	Sumers et al. [4]	This survey
Primary scope	Memory mechanisms of LLM agents	Cognitive architectures for agents	Evidence status of agent memory
Organizing logic	What, where, when to store	Memory within action and decision loops	Capacity, consistency, control
Failure evidence	Selected examples	Conceptual	Quantified benchmarks, editing and extraction studies
Governance treatment	Brief	Absent	Unlearning, privacy regulation, auditable lifecycles

Our framework poses three questions of any memory claim. Capacity: how much does the

system actually retain and use—measured, not advertised? Consistency: does what it retains remain correct when the world changes, when retrieved content conflicts with parametric knowledge, or when one stored fact implies another that was never updated? Control: can what is remembered be inspected, corrected, and removed on demand, by operators and users with legitimate authority? The three are ordered: capacity without consistency is confident error at scale, and neither matters if memory cannot be governed.

The ordering deserves defense, because the field's incentives run the other way. Capacity is the glamorous property: it sells products, anchors headlines, and is the dimension along which systems are most often compared. But a memory system's failures grade sharply in severity as one moves down the list. A capacity failure—forgetting a preference, missing a detail—degrades usefulness. A consistency failure—acting on a corrected address that was "updated" in the store but not in the agent's behavior, re-asserting a rescinded instruction—creates harm with a paper trail that says otherwise: the system's record shows the right thing while its behavior does the wrong one, which is the worst evidentiary position a governed system can occupy. A control failure—no verified way to remove what a user demanded removed—converts both prior failures from incidents into liabilities, because there is no remediation after the fact. The order is therefore also a priority order for governance: the sections that follow assess the evidence for each gap in turn, and the evidence, as will be seen, degrades precisely down the list—the capacity gap is the best measured, the control gap hardly measured at all. Section 2 reviews the architectural generations the field has already built; Section 3 assesses the evidence for each gap; Section 4 draws the governance consequences; Section 5 concludes.

2. The Agent Memory Stack: Three Architectural Generations

2.1. Prototypes: Operating Systems and Memory Streams

The founding metaphor of agent memory is the operating system. MemGPT reframes the context window as main memory and external storage as disk, and has the LLM itself interrupt its generation to page information between the two—searching archives, revising its own working contents, deciding what deserves scarce in-context space [5]. The significance of the design is less the particular mechanism than the transfer of authority: memory management ceases to be a fixed developer-imposed pipeline and becomes behavior of the model at run time. Generative Agents, simultaneously, demonstrated what a rich memory stream makes possible: every experience is recorded as an observation; retrieval weights recency, importance, and relevance; and a periodic reflection operation synthesizes lower-level memories into higher-

level inferences that are themselves stored and retrievable [6]. The simulated small town built on this machinery exhibited emergent social behavior—one agent organizing a Valentine's Day party that other agents attended—that no single prompt produced [6].

Together the two prototypes defined the agenda that later systems industrialize: memory as an explicit, self-managed object rather than a passive context buffer. They also fixed a pattern that Section 3 returns to: both systems were validated primarily by demonstration and simulation—emergent behavior, human evaluations of believability—rather than by instrumented measurement of what the memory layer itself retained, weighted correctly or forgot on schedule. The architectural claims ran ahead of the evidential ones from the start, and the production generation inherited the asymmetry along with the machinery.

Each prototype also contributed a mechanism that later systems treat as standard, and the mechanisms carry governance weight their papers did not claim for them. MemGPT's paging established that the model itself can decide what enters scarce memory—an architecture in which retention policy is emergent model behavior rather than developer specification [5]. Generative Agents' reflection established that the store can rewrite itself, synthesizing observation-level memories into inferences that enter the store as first-class entries with their own retrieval weight [6]. Both mechanisms are why memory contents are hard to predict from the outside: what an agent remembers is partly a function of its own judgments at run time, which means an auditor cannot enumerate the store by replaying the inputs—the store is a product of decisions, not a log. The distinction between a log (append-only, replayable, complete by construction) and a memory (curated, synthesized, self-modifying) is the prototypes' deepest legacy, and Section 3's control gap is in large part the absence of any evaluated answer to the question it poses: when the record is a distillation, what does it even mean to inspect or correct it? [5,6]

The prototypes also explain why the memory literature's vocabulary is so unevenly operationalized. Terms that later papers use as engineering nouns—episodes, reflection, importance, forgetting—entered the field as design metaphors from the two systems, carrying their simulation-era connotations with them; MemoryBank's forgetting curve is Ebbinghaus applied as a decay function rather than as psychology, and A-MEM's "evolved context" is Generative Agents' reflection generalized to link propagation. Metaphor inheritance is not a defect—every engineering field borrows its early vocabulary—but it does mean the field's central terms name mechanisms before they name measured properties: no paper in this survey's corpus that says "forgetting" reports a forgetting rate, and none that says "reflection" reports a

reflection accuracy. The vocabulary gap is the gaps of Section 3 in miniature, visible already at the prototypes' origin: a system vocabulary that outruns its measurement vocabulary at birth tends to keep the lead, and the production generation's benchmark suites—valuable as they are—did not close it [5,6].

2.2. Production Systems: Memory as a Service

The second generation packages memory as deployable infrastructure. Mem0 extracts candidate memories from conversations, consolidates them against existing entries—adding, updating, or deleting—and serves them back through a retrieval pipeline, reporting on the LOCOMO long-conversation benchmark both accuracy exceeding full-context baselines and large savings in latency and token cost [7]. The comparison matters: LOCOMO itself evaluates agents on conversations hundreds of turns long and finds that faithfully answering requires aggregating evidence scattered across them, a capability in-memory pipelines are still failing at differentially across information types [8]. Memory-as-a-service, in other words, is already being benchmarked against the hardest thing memory must do—keeping a coherent account of a long shared history—and the benchmarking is genuinely adversarial. A-MEM pushes the agentic organization of memory further: when a new note arrives, the system dynamically generates links to related prior notes and propagates an evolved contextual description backward through the network, so the memory structure itself keeps reorganizing as the agent's history grows [9]. MemoryBank approaches the same lifecycle from the temporal side, attaching an Ebbinghaus-style forgetting curve to stored memories so that unreviewed content decays and a long-term companion persona reflects the natural attrition of an organic relationship [10].

The production generation's distinctive commitment is lifecycle management—write, consolidate, link, decay, delete—as an explicit service boundary. What it has not yet produced is a shared evidential standard for the boundary: each system is evaluated on partly different suites, and the persuasive force of the numbers lies in comparisons against context baselines rather than against the retention, correctness, and removability guarantees that the lifecycle language implies.

Two details of the production generation's own evidence qualify its headline numbers, and both matter for what the systems can be claimed to do. First, the strongest benchmark results are comparative: Mem0's reported gains are measured against full-context and baseline-pipeline configurations on LOCOMO, which establishes that the pipeline beats the alternatives tested—not that it retains, stays consistent with, or can verifiably forget any particular fact category [7].

Second, the benchmark that anchors the comparison evaluates memory at a difficulty deliberately beyond current systems: LOCOMO's conversations run to hundreds of turns with evidence deliberately scattered, and its own analysis shows systems failing differentially across information types—multimodal and temporal items hardest—which means even the favorable comparisons sit on a floor of substantial absolute error [8]. Neither qualification is an accusation; both are what honest benchmark evidence looks like. The governance reading is about the gap between the evidence and the marketing: a pipeline whose measured behavior includes substantial failure on temporal and scattered evidence is being sold, in lifecycle language, as a system that keeps the record current. The lifecycle vocabulary arrived before the lifecycle guarantees did [7,8].

2.3. The Retrieval Frontier: Indexing Before Storing

Beneath the system papers lies a retrieval lineage. Retrieval-augmented generation established that grounding generation in documents fetched at query time is the workable alternative to cramming knowledge into parameters or context [11]. HippoRAG imports the neurobiology of episodic memory to improve that grounding: an offline step builds a knowledge graph from the corpus as an artificial hippocampal index, and retrieval runs a personalized PageRank over it, mimicking the pattern-completion dynamics of hippocampal indexing theory; the result is markedly stronger multi-hop question answering than dense retrieval alone, at a fraction of the cost of iterative agentic retrieval [12].

Read across its three generations, the stack shows a consistent trajectory: memory decisions migrate from the developer's blueprint into run-time model behavior—paging, reflecting, linking, forgetting—and the store itself becomes richer and more structurally self-modifying. Table 2 summarizes the generations.

The retrieval lineage's governance property is worth stating precisely, because it is the stack's best and rarely advertised. A retrieval-indexed memory has the one property regulators and auditors know how to work with: enumerability. The store's contents can be listed; a given memory is either in the index or not; a deletion is an index operation whose effect is complete and immediate; and the provenance of any retrieved item is, in principle, a field in its record. RAG established the architecture [11], and HippoRAG's contribution—the graph index with spreading-activation retrieval—compounds the property rather than departing from it: a graph over notes is still an enumerable structure, and its personalized-PageRank retrieval is a deterministic function of the graph, so multi-hop influence is traceable in a way that parameter-internal influence is not [12]. When later sections document the capacity, consistency, and

control gaps, the retrieval frontier is where the correctives are cheapest to install—and the production systems that emphasize neural consolidation over indexed structure are, knowingly or not, trading away enumerability, the very property that governance will eventually ask them to produce. The stack's most sophisticated corner is thus its least auditable, by design rather than by accident. The trajectory is the field's strength and its exposure: the same autonomy that makes memory useful makes its contents harder to predict from the outside, which is why the gaps of Section 3 fall where they do.

Table 2. Three generations of agent memory systems.

Generation	Representative systems	Organizing metaphor	Who manages memory	Evidence style
Prototypes (2023)	MemGPT [5], Generative Agents [6]	OS paging; memory stream	The model, at run time	Simulation, human evaluation
Production (2024–25)	Mem0 [7], A-MEM [9], MemoryBank [10]	Memory as a service	Dedicated pipeline	Benchmark suites, e.g., LOCOMO [8]
Retrieval frontier (2024)	RAG [11], HippoRAG [12]	Hippocampal indexing	Offline index + propagation	Multi-hop QA accuracy

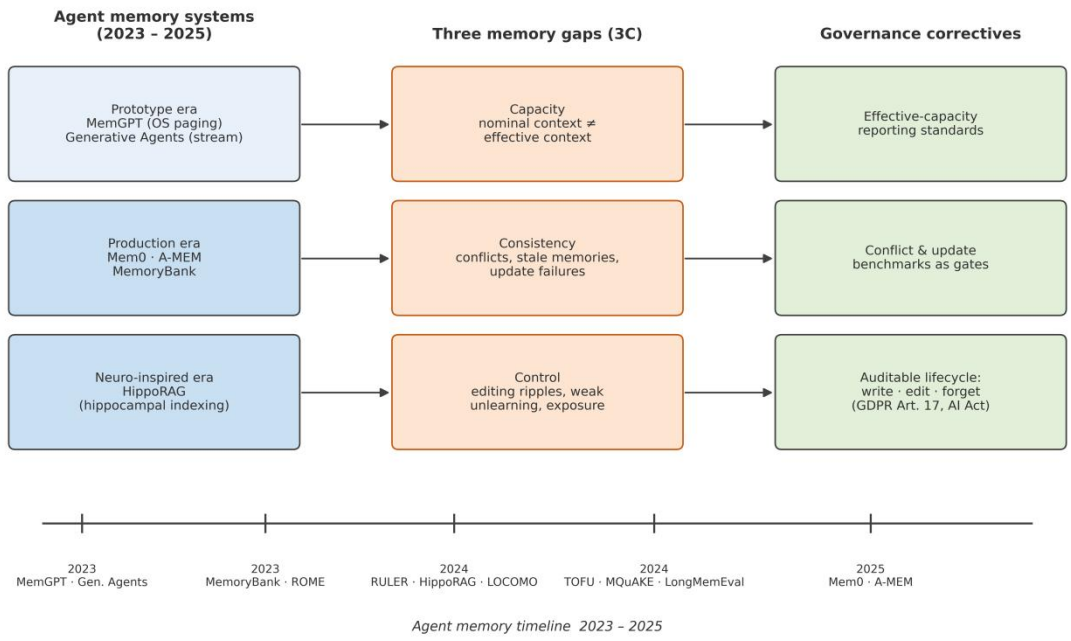

Figure 1. The three memory gaps of agent memory systems. Left: three architectural generations of the memory stack. Middle: the capacity, consistency, and control gaps separating built systems from evidenced memory. Right: the governance correctives each gap demands. Bottom: the field's timeline, 2023–2025.

3. Three Gaps: Capacity, Consistency, Control

The systems of Section 2 exist and are deployed. This section asks what is actually known about them, gap by gap. The organizing claim is that each gap is not a missing feature but a mismatch between what has been built and what has been measured—and that the mismatch is widest precisely where deployment pressure is strongest.

3.1. The Capacity Gap: Nominal Is Not Effective

The capacity gap is metrological: nominal context length is a configuration, effective context length is a measurement, and the two diverge systematically. RULER operationalizes the difference with a task ladder whose rungs demand progressively deeper use of the context—from simple needle retrieval through variable tracking to multi-hop aggregation across many distributed items—and shows that a model's effective length is not one number but a function of task complexity, with models that hold up on retrieval collapsing well before their claimed limits on aggregation [1]. The practical reading is direct: a context window advertised at 128,000 tokens is not a promise that 128,000 tokens are usable; on the harder rungs, measured capacity for several models falls to a fraction of the claim [1]. Positional structure further degrades the usable fraction, since mid-context material is systematically underweighted even when retrievable in principle [2].

Economics closes the trap from the other side. Attention cost grows with context, so retention by stuffing is priced per token and per query—an economic boundary independent of any benchmark. The memory stack of Section 2 is therefore not a temporary workaround awaiting bigger windows; it is the rational response to a window that is expensive, transient, and only fractionally effective. But the same reasoning transfers the burden: once memory lives in an external store, the system's true capacity is whatever the combined pipeline retains and uses, and that quantity—external store, retrieval quality, in-window positional effects compounded—is precisely what no headline specification measures. Capacity, in short, is currently claimed in tokens and evidenced in benchmarks; the gap between the two currencies is the first thing an operator must audit.

The measurement details sharpen the point from abstract to actionable. RULER's task ladder is public and parameterized, which means an operator can locate where on the ladder their deployed configuration begins to fail—and the interesting finding is that the failure point is not a fixed token count but moves with task depth: the same model that retrieves a needle at 90,000 tokens may fail to aggregate across 40 needles at 30,000 [1]. The lost-in-the-middle curve adds

a positional dimension to the audit: it is not only how much is in the window but where—the same information placed in the middle of a long context is substantially less likely to be used than when placed at the edges, a U-shaped usability surface inside every nominally uniform window [2]. For agent memory design the two results compose into a siting problem: an external store that stuffs retrieved history into the middle of a long prompt is placing its most valuable contents in the least reliable real estate of the window, and an operator who sizes the retrieval budget against the nominal window rather than the effective one is systematically overfilling the prompt. Capacity measurement, in other words, is not merely a benchmark formality—it dictates concrete retrieval-configuration decisions that production systems currently make against the wrong number [1,2].

3.2. The Consistency Gap: Conflicts, Updates, and Wrong Recency

A memory that grows across sessions must survive a changing world, and here the evidence is most concerning. Start with conflict. Xie et al. show that when contextual evidence contradicts parametric knowledge, model behavior is unstable in a characteristic way: performance on knowledge-conflict tasks drops compared to clean settings, and models frequently persist in stale parametric answers even against explicit contextual evidence [13]. Agent memory industrializes this hazard, because external stores accumulate exactly the kind of plausible, fluent, occasionally outdated content that competes with the model's prior beliefs at every retrieval. The benchmark evidence concurs from the evaluation side. LongMemEval decomposes long-term interactive memory into five abilities—information extraction, multi-session reasoning, temporal reasoning, knowledge updating, and abstention—and finds knowledge updating and temporal reasoning among the weakest for commercial chat assistants, with accuracy dropping substantially on long-context variants of its suite [14]. LOCOMO reaches a compatible verdict from conversation scale: across hundreds of turns, the systems it evaluates struggle to correctly aggregate and recall information distributed across a shared history, with failures concentrated on exactly the multi-session and temporal categories that memory exists to serve [8]. Memory, in other words, is weakest at its defining task—keeping the record current.

LongMemEval's decomposition is the most diagnostic instrument in the corpus because it separates the failure modes that "memory" blurs together, and the separation changes what one would fix. Its five abilities—information extraction, multi-session reasoning, temporal reasoning, knowledge updating, and abstention—are scored separately, and the published pattern is that knowledge updating (revising a stored fact when the user contradicts it) and

temporal reasoning (ordering and dating remembered events correctly) sit at the bottom of commercial systems' performance [14]. Abstention is the most consequential of the five and the least discussed: knowing that the store does not contain the answer, and saying so, is what prevents an agent from confabulating a plausible detail on top of an empty retrieval—and it, too, degrades as sessions lengthen [14]. A system designer reading the decomposition gets an actionable map: temporal failures call for timestamp discipline in the store and schema, updating failures call for the propagation mechanics of the editing literature, abstention failures call for calibrated retrieval confidence rather than more capacity [14]. The single-number memory scores that circulate in marketing hide exactly the distinctions that determine which engineering fix applies—which is why the benchmark's decomposition, not any aggregate, is the consistency gap's real measurement [14].

Editing studies sharpen the point into a structural claim. Model-editing methods can rewrite a stored association in place—rank-one updates locate and modify the weight positions that mediate a factual recall [15]—and yet MQuAKE shows that such edits fail to propagate: correcting one fact leaves the model asserting the old one whenever the answer requires traversing it through a multi-hop chain, so the correction is present in the store and absent from the reasoning [16]. For agent memory the implication is uncomfortable: consistency is not a property of individual writes but of the inference graph over all of them, and the current tooling treats the store as a collection of independent notes. Until edit operations carry their ripple structure—until updating one memory invalidates what it entailed—consistency will be measured as an afterthought rather than guaranteed by construction.

The mechanics behind the propagation failure explain why it is structural rather than a tuning problem. ROME's causal-tracing procedure localizes a single factual association to specific mid-layer positions and rewrites them with a rank-one update—an operation whose beauty is its narrowness: it changes one association's storage while barely disturbing neighboring weights [15]. That narrowness is exactly what MQuAKE exposes on the reasoning side: a multi-hop answer ("Where was the wife of the director of film X born?") requires the corrected fact to participate in a chain, and the un-traversed edges of the chain still carry the pre-edit association [16]. Agent memory reproduces both halves at the system level: the write-side narrowness lives in memory pipelines that treat entries as independent rows, and the read-side propagation failure appears whenever generation must compose several stored memories—precisely the multi-session aggregation that LOCOMO and LongMemEval identify as the hardest categories [8,14]. Consistency, on this evidence, is a graph property being managed with list operations: until

write operations declare and propagate their entailments, no amount of per-entry correctness checking will make the store consistent, because the inconsistency lives between the entries, not inside any of them [15,16].

3.3. The Control Gap: Editing, Forgetting, and Exposure

The control gap concerns authority over the store: who may write, correct, and—critically—remove. Machine unlearning is the formal counterpart of the human right to be forgotten, and its current state is fragile. TOFU frames unlearning on fictitious author profiles, where the ground truth of what "forgotten" means is knowable, and finds that existing techniques preserve model utility well while largely failing to truly remove the targeted knowledge—and that the gap widens drastically as the amount to be forgotten grows [17]. Extraction results give the removal failure its privacy stakes: Carlini et al. demonstrate that training data, including personally identifiable information, can be recovered from large models by black-box queries, establishing memorization as an exploitable surface rather than a theoretical concern [18]. An external memory store reproduces that surface in plainer form—its contents are readable, copyable, and subject to subpoena or breach—while compounding it with durability: what a user told an agent yesterday may persist in a vendor's store long after the conversation itself is gone.

First-party practice shows both responsiveness and asymmetry. OpenAI's documentation for ChatGPT memory describes reference memories that users can view, correct, and delete individually, and sessions whose memories are cleared—but the decisions of what to write in the first place remain interior to the product, opaque to the user whose history supplies them [19]. The asymmetry is the gap in miniature: removal has a user-facing control surface, while writing and retention-policy do not. No benchmark in the survey's corpus measures write justification, retention horizon, or deletion completeness as first-class properties of a memory system. Control, at present, is a set of scattered affordances rather than an evaluated property.

The extraction evidence supplies the control gap's adversarial boundary, and it is worth stating in the store-specific form. Carlini et al.'s methodology— inducing the model to diverge from its fluency prior so that memorized, high-perplexity training text surfaces—shows that "the model does not reveal it on normal queries" is not a privacy property but a latency property: the data stays in the artifact, and the artifact's resistance is empirical, not structural [18]. An external memory store makes the corresponding point without any attack: its contents are directly readable, exportable, and subject to legal process, so the question for the store is never whether the data can be extracted but who may read the store and on what showing. That inverts

the security posture relative to parameters—in weights, extraction is the attack; in the store, access is the attack—and governance instruments exist for the second case. What does not yet exist is the verification layer either way: whether data was truly removed from weights or truly deleted from the store, the operator's claim currently cannot be checked by the data subject, and TOFU's probing protocol is the only published template for what such a check would look like [17]. Control, read adversarially, is therefore two problems—access control on stores, extraction resistance on parameters—plus one shared missing piece: verifiable deletion, which neither the store operators nor the model providers currently offer evidence for [17,18].

4. Governance: Memory as an Auditable Record

4.1. Evaluation as Infrastructure

The first corrective is already under construction: the benchmark families of Section 3. If capacity must be reported as effective length under a task ladder, consistency as update and conflict performance, and control as deletion and abstention behavior, then LongMemEval's five-ability decomposition and LOCOMO's long-conversation protocol are the beginnings of a metrology for memory [8,14]. What benchmarks do well is make claims comparable and regressions visible; what they cannot do is certify behavior on future inputs. The governance reading is therefore modest and concrete: procurement and deployment decisions should treat effective-capacity and update-failure figures as gating information, exactly as throughput figures gate other infrastructure purchases.

There is a precedent for benchmarks becoming governance artifacts, and it explains why the modest reading is not a small one. Seat-belt crash ratings, university rankings, and financial stress tests all began as measurement exercises whose published numbers acquired regulatory and market force—the measurement became the governance, without any legislature involved. The memory benchmarks are positioned identically: RULER's task ladder, LongMemEval's five abilities, and LOCOMO's long-conversation protocol are public, parameterized, and comparable across systems [1,8,14], which is precisely the property that lets a procurement officer write "effective capacity at aggregation depth $X \geq Y$" into a contract, or a regulator write "update-failure rate below Z" into a condition of deployment. The benchmarks' authors did not design them for this role—RULER asks what a model's real context size is; LOCOMO asks how agents fare over very long conversations [1,8]—but governance does not require the design intent, only the measurement's public comparability. The gap that remains is coverage: no benchmark in the corpus measures write-side justification or deletion completeness as first-

class scores, so the gatekeeping that benchmarks enable currently covers the capacity and consistency gaps and skips the control gap [8,14].

4.2. The Regulatory Surface

The legal surface is closer than much of the technical debate assumes. The GDPR grants data subjects a right to erasure of personal data, with narrow exemptions, and an agent memory store containing user history is personal-data processing in an unusually literal form: persistent, structured, attributable to a person [20]. The EU AI Act overlays risk-tiered obligations on providers and deployers, including logging and record-keeping duties for high-risk systems—obligations that presuppose the kind of inspectable memory whose absence Section 3 documented [21]. The combined reading is that European law already treats what agents remember as governed data; what is missing is the engineering correspondence—memory systems designed so that erasure requests, retention limits, and audit queries are operations with verified outcomes rather than best-effort deletions.

The regulatory structure maps onto the three gaps with uncomfortable precision, which is the survey's framework paying off analytically. The GDPR's storage-limitation principle binds what agents retain—a capacity obligation stated in law but unimplemented in any memory system's design or evaluation, since no system in Section 2 measures or even declares a retention horizon [20]. The accuracy principle of the same regulation—personal data must be kept correct and up to date—binds what agents believe: an agent acting on a stale address or a rescinded preference is processing inaccurate personal data, whatever the store's own entry says, which makes the consistency gap a compliance question rather than a quality question [20]. And the AI Act's record-keeping and logging duties for high-risk systems presuppose that what the system did and why can be reconstructed after the fact—an obligation the self-modifying stores of Section 2.2 make structurally difficult, since consolidation and reflection rewrite the very traces the logs are supposed to preserve [21]. Three legal obligations, three unimplemented engineering properties: the mapping is not a coincidence but the point—European law was written for records, and agent memory is now a record-keeping technology that has not yet accepted the description [20,21].

4.3. The Auditable Memory Lifecycle

The synthesis is to treat memory as a record with a lifecycle, and to demand for each stage the properties governance requires of records elsewhere. Write with justification: entries carry their provenance and the evidence that warranted retention. Maintain with consistency: updates

propagate to entailed content, in the direction MQuAKE showed current editing fails [16]. Forget with verification: deletion is validated by probing, in the spirit of TOFU's ground-truthed unlearning [17], and extraction attacks [18] define the adversary the verification must defeat. Answer with traceability: every retrieved memory carries an audit trail to its origin. None of these properties is exotic; all of them are measurable; and the three gaps of this survey are, in governance terms, precisely their current deficits.

A concrete deployment sketch shows the lifecycle is implementable with current parts, which matters because "auditable memory" risks reading as aspiration. On the retrieval frontier's indexed stores [12], write-side justification is a schema field; propagation on update is a graph query over entailment edges—expensive, but bounded and checkable in a way parameter-side ripples are not; deletion is an index operation, and its verification can reuse TOFU-style probing against the store rather than the weights, where ground truth is knowable by construction [17]. The hard stages concentrate exactly where the stack is least enumerable: consolidated and self-synthesized memories, whose provenance the prototypes' reflection operations deliberately blur [6], and the parametric residue of what models bring to every read. The engineering conclusion is therefore a siting decision, not a research program: put governed knowledge in the enumerable part of the stack where the lifecycle properties are cheap, and treat the non-enumerable parts as unverified by policy—subordinate to the indexed store for anything regulation touches—rather than as coequal memory. Vendors will resist the demotion; the evidence of Section 3 says the demotion is what "auditable" costs [6,12,17].

5. Conclusion

This survey organized agent memory along three gaps and reached three findings. First, the architectural side has consolidated across three generations—OS-inspired prototypes, production memory services, and neurobiologically inspired indexing—with memory management migrating steadily into run-time model behavior [5-7,12]. Second, the evidential side lags the architectural one at all three of our questions: effective capacity falls short of nominal capacity under task load [1,2]; consistency fails exactly where memory's defining duty lies, in updates and temporal grounding [8,13,14,16]; and control remains a scattered set of affordances without a measured guarantee of removal [17-19]. Third, the governance apparatus needed to close the gap is partially in place—benchmarks provide a metrology [8,14], and European law already asserts authority over what agents remember [20,21]—but no evaluated lifecycle yet connects the two.

The framework travels beyond chat agents. Any system that accumulates operational history—recommendation platforms, autonomous-process controllers, institutional knowledge bases—faces the same triad of claims outrunning evidence. For agent memory specifically, the highest-value open problems follow the gaps: reporting standards that state effective capacity under task load rather than token ceilings; editing and update operations with verified ripple propagation; unlearning whose completeness is probed rather than presumed; and write-side transparency commensurate with the deletion controls that first-party products already ship. Memory is becoming the load-bearing component of agentic systems; making its claims auditable is how the component earns the load.

The near-term trajectory the evidence supports is a division of labor rather than a single convergence. Inside the window, the capacity benchmarks will keep measuring effective retention as windows grow, and the positional economics will keep making external memory the rational default for anything durable [1,2]. Outside the window, the indexed, enumerable corner of the stack will absorb the governed workloads first—because it is where the lifecycle properties of Section 4 are cheapest to implement and verify—while consolidated, self-modifying memory earns trust exactly as fast as its write, propagation, and deletion claims become probeable [12,17]. None of this requires a scientific breakthrough; it requires the field to treat memory the way databases were eventually forced to treat durability—as a property with a definition, a test, and a failure report—rather than the way it is treated today, as a feature with a demo. The three gaps name the distance between those two treatments; closing them is, in the most literal sense, a matter of record.

References

[1] Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., & Ginsburg, B. (2024). RULER: What's the real context size of your long-context language models? arXiv preprint arXiv:2404.06654.

[2] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12, 157-173.

[3] [3] Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., Zhu, J., Dong, Z., & Wen, J.-R. (2024). A survey on the memory mechanism of large language model based agents. arXiv preprint arXiv:2404.13501.

[4] Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive architectures for language agents. arXiv preprint arXiv:2309.02427.

[5] Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). MemGPT: Towards LLMs as operating systems. arXiv preprint arXiv:2310.08560.

[6] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. ACM.

[7] Chhikara, P., Khant, D., Aryan, S., Singh, T., & Yadav, D. (2025). Mem0: Building production-ready AI agents with scalable long-term memory. arXiv preprint arXiv:2504.19413.
[8] Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024). Evaluating very long-term conversational memory of LLM agents. arXiv preprint arXiv:2402.17753.
[9] Xu, W., Liang, Z., Mei, K., Gao, H., Tan, J., & Zhang, Y. (2025). A-MEM: Agentic memory for LLM agents. arXiv preprint arXiv:2502.12110.
[10] Zhong, W., Guo, L., Gao, Q., Ye, H., & Wang, Y. (2024). MemoryBank: Enhancing large language models with long-term memory. arXiv preprint arXiv:2305.10250.
[11] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33.
[12] Jiménez Gutiérrez, B., Shu, Y., Gu, Y., Yasunaga, M., & Su, Y. (2024). HippoRAG: Neurobiologically inspired long-term memory for large language models. In Advances in Neural Information Processing Systems 37.
[13] Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In The Twelfth International Conference on Learning Representations (ICLR).
[14] Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., & Yu, D. (2024). LongMemEval: Benchmarking chat assistants on long-term interactive memory. arXiv preprint arXiv:2410.10813.
[15] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems 35.
[16] Zhong, Z., Wu, Z., Manning, C. D., Potts, C., & Chen, D. (2023). MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
[17] Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). TOFU: A task of fictitious unlearning for LLMs. In First Conference on Language Modeling (COLM).
[18] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633-2650).
[19] OpenAI. (2024). Memory FAQ. OpenAI Help Center. https://help.openai.com
[20] European Parliament and Council. (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). Official Journal of the European Union, L 119.
[21] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

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

From Corpus to Model: Governing the Knowledge Supply Chain of Large Language Models

Ruyue Yang, Xinyu Song*
Geely University of China, ChengDu, China, 641423

Received: August 22 2026

Revised: August 30, 2026

Accepted: August 30, 2026

Published online: August 30, 2026

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

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

Abstract. Large language models acquire most of their knowledge from sources—web corpora, licensed archives, curated datasets—whose governance traditionally depends on a capability that parametric learning destroys: the ability to take knowledge back. This survey reviews 18 core sources (2018–2024) through an original knowledge supply chain framework that traces knowledge from upstream sourcing and licensing, through midstream injection into model parameters or retrieval indexes, to downstream use and recall, with a governance control point attached to each stage; reported findings are quoted only from primary sources and court records. Three findings emerge. Upstream, the data foundation is weakly documented: large-scale audits find that most open datasets do not convey their own licensing terms, and web-scale corpora contain removed and toxic content at unrecorded rates. Midstream, the four knowledge-injection channels differ enormously in reversibility—retrieval is fully reversible, model editing partially so, pre-training effectively not at all—yet systems treat them interchangeably. Downstream, removal is the weakest control: machine unlearning remains unverified at scale exactly as litigation begins to demand it. We map documentation standards, reversibility grading, and verified deletion onto the chain's stages as actionable control points.

Keywords: *data provenance; machine unlearning; retrieval-augmented generation; licensing; dataset documentation; copyright*

1. Introduction

1.1. The Lawsuit That Exposed a Missing Capability

On December 27, 2023, The New York Times Company sued Microsoft and OpenAI in the

Southern District of New York, alleging that millions of its articles were used to train large language models without permission and that the models can reproduce near-verbatim stretches of its journalism [The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y.)] [1]. Whatever its outcome, the complaint exposes a structural asymmetry that no verdict can repair. In ordinary data governance, inclusion is reversible: a record entered without authorization can be deleted, and deletion propagates to every system that respects the source of truth. In a trained language model, the copy that matters is not a record but a distributed change in billions of parameters, entangled with everything else the model learned from the same corpus. There is no query that removes it, no index to drop it from, and no rollback to a pre-training state that would not also erase everything else. Knowledge, once injected into parameters, has left every regime in which "remove it" is a meaningful instruction.

The procedural posture of the case amplifies the asymmetry rather than softening it. The complaint's requested relief reportedly includes destruction of models trained on infringing copies—an equitable remedy that, in ordinary product governance, corresponds to a full recall of a manufactured good, an event rare and catastrophic enough that entire insurance and inspection industries exist to prevent it [1]. Whatever disposition the court reaches, the industry's behavior since filing suggests it has accepted the underlying diagnosis: licensing negotiations between model developers and content owners have proliferated, opt-out registries have become negotiating instruments, and disclosure of training-data composition has moved from an academic demand to a term of commerce. These are upstream and contractual responses to a defect the industry has not learned to remediate once knowledge is already in the weights—which is precisely why the case matters beyond its parties: it documents, in the most authoritative forum available, that the governance toolkit currently ends where training begins.

This survey treats that asymmetry as the defining governance problem of large language models and organizes the field's evidence around it. The organizing image is a supply chain: knowledge is sourced upstream, injected midstream, used downstream, and—if governance requires—recalled, each stage with its own documented failure modes and its own candidate control points.

1.2. Scope and Method

We focus on the governance of knowledge in large language models—where it comes from, how it enters, how it stays correct, how it leaves—rather than on the deployment governance of model applications (bias in hiring, content moderation), which is treated extensively elsewhere. We searched arXiv and major machine-learning venues for 2018–2024 primary

sources, complemented by court records and European Union legislative texts. Sources were retained if they (i) audit or document the provenance and licensing of training data, (ii) propose or evaluate a mechanism that injects, corrects, or removes knowledge, or (iii) constitute primary governance instruments. Purely positional commentary on AI copyright was excluded; the survey quotes the litigation only for its documented factual claims. This procedure yielded the 18 core sources organized below.

The verification protocol deserves a sentence of its own, because governance writing inherits the defects it audits unless it disciplines itself. Every empirical claim in what follows is quoted from the audited source itself—benchmark papers for benchmark numbers, audits for audit statistics, court records for litigation facts—rather than from commentary about them; each source's author list, title, and venue were checked against its primary record before inclusion; and where a citation's provenance could not be verified with confidence, the source was dropped rather than cited on recollection. Two consequences follow. First, the survey's evidential base is deliberately narrower than the field's literature—it contains no secondary summaries of failure rates, no numbers remembered from other surveys—and its figures can be traced to a primary document in one step. Second, and more usefully for the reader, the verification burden itself models the survey's subject: the demand that a claim carry its provenance is exactly the upstream record this survey will argue the model-knowledge supply chain lacks. The survey asks a standard of the field's knowledge claims that it attempts to meet for its own. The survey deliberately overlaps with, but is distinct from, agent-memory surveys: memory systems govern what an agent writes down between sessions, while this survey governs what a model is—its parameters, corpora, and retrieval sources.

1.3. Positioning and the Supply Chain Framework

Three prior syntheses each illuminate one region of the problem without spanning it. Bommasani et al. introduced the foundation-model frame and catalogued the societal stakes of models trained on undifferentiated data, but as a broad agenda document it does not organize technical evidence by governance stage [2]. Ji et al. survey hallucination in natural language generation, giving the field its standard intrinsic/extrinsic taxonomy of unfaithful output—a midstream-to-downstream maintenance failure treated in isolation from where the knowledge came from [3]. The Data Provenance Initiative audits dataset licensing and attribution at scale—an upstream audit that stops at the dataset boundary and says nothing about what training does with what it documents [4]. Table 1 contrasts the three with this survey.

Table 1. Coverage comparison between prior syntheses and this work.

Dimension	Bommasani et al. [2]	Ji et al. [3]	Data Provenance Initiative [4]	This survey
Primary scope	Foundation-model agenda	Hallucination taxonomy	Dataset licensing audit	Knowledge lifecycle governance
Chain coverage	Broad, non-staged	Midstream–downstream	Upstream only	Upstream through recall
Failure evidence	Framing arguments	Taxonomy + mitigations	License audit statistics	Quantified, all stages
Governance output	Research agenda	Mitigation methods	Compliance tooling	Control points per stage

The framework's central term deserves definition before it is used. A governance control point is a stage-attached intervention at which three properties can be established: attribution (the intervention names the knowledge it governs—this corpus, this channel, this output), verification (some procedure can check that the intervention did what it claims—audit, probing, record inspection), and enforceability (an actor with authority—court, regulator, procurer, operator—can compel or reward compliance). Mature supply-chain regulation, from food safety to electronics, is built from exactly such points: documented inputs, graded intermediates, traceable outputs, and recall authority at each stage. Read against that template, the model-knowledge chain fails not at any single point but in a specific pattern: attribution is weakest upstream (untraceable corpora), verification is weakest at recall (unproven unlearning), and enforceability, which exists in law, currently attaches to nothing it can grip. The survey's analytical claim is that these are engineering specifications, not philosophical puzzles—each section will name what a working control point at its stage would measure.

The supply chain framework gives each stage a control question. Upstream: can the provenance and permission of every source be established? Midstream: through which channel does knowledge enter, and how reversible is that channel? Downstream: does output remain faithful to sources that may themselves have changed? Recall: can knowledge be removed when law or correction demands it? Section 2 takes the stages in order; the framework's payoff is that remedies attach to stages—a datasheet cannot fix a ripple effect, and an unlearning benchmark cannot fix an unlicensed corpus. Figure 1 previews the chain, its documented failures, and its control points.

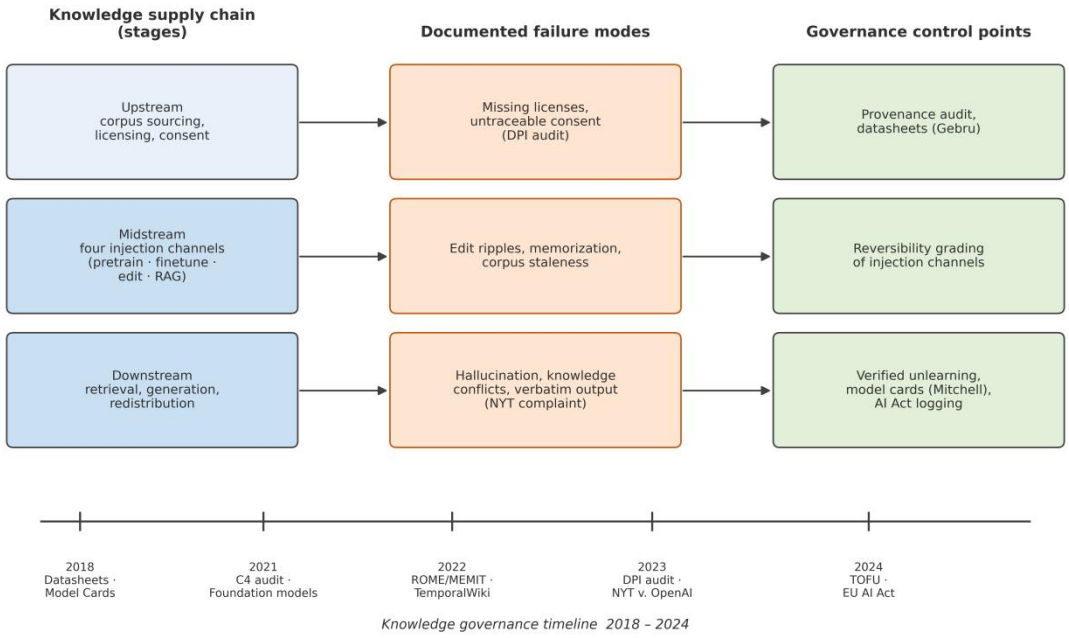

Figure 1. The knowledge supply chain of large language models. Left: the three lifecycle stages plus recall. Middle: documented failure modes at each stage. Right: governance control points. Bottom: the field's timeline, 2018–2024.

2. The Upstream: Provenance Without Permission

The chain begins with corpora that are, by the field's own audits, weakly governed. Dodge et al. supplied the first systematic documentation of C4, the Colossal Clean Crawled Corpus behind many prominent models, and found a dataset assembled by heuristic filtering rules with essentially no provenance discipline: pages disappear after scraping at unrecorded rates, filtering silently drops whole registers of text, and the surviving content skews toward particular dialects and publishers in ways no one had measured [5]. The documentation gap is not cosmetic; it means that when a model trained on such a corpus later emits a copyrighted or toxic passage, the data to trace where that passage came from does not exist in any usable form.

The licensing picture is weaker still. The Data Provenance Initiative audited nearly 2,000 widely used open datasets and their thousands of source documents, and found that more than half of them fail to convey license information usable in practice—terms are missing, misdescribed, or attached to sources they do not actually govern [4]. In supply-chain terms, most of the field's raw material arrives without a paper trail: not necessarily acquired wrongfully, but unauditable, which for governance purposes is the operative defect. An unauditable input cannot be excluded ex ante, priced, or recalled ex post; it can only be litigated after the fact, which is precisely the position the NYT complaint [1] occupies. The upstream control point is therefore documentary before it is legal—establishing what entered, under what terms—

because every downstream control presupposes that record.

The texture of the C4 audit is worth conveying, because it shows that the documentation failure is systematic rather than a matter of a few negligent sources. Dodge et al.'s methodology was itself an act of reconstruction: they manually inspected samples of the corpus and traced documents back to their URLs, classifying what remains of each origin page. A large majority of the traced documents no longer resolve to their original content—the pages were deleted, moved, or changed after crawling—so the corpus now contains text whose source context is unrecoverable from the live web, and the audit's classification of content type and quality had to proceed from the scraped snapshots alone [5]. Two findings carry particular governance weight. Filtering, the audit shows, was where most silent loss occurred: the corpus's cleaning heuristics, designed to remove boilerplate and gibberish, differentially removed entire registers of text, with the surviving document distribution skewed toward a narrow band of publishers and registers—and the corpus's documented racial, gender, and toxicity profiles follow that skew rather than any property of the web itself [5]. In supply-chain terms, the cleaning step was a processing operation that changed the product's composition without leaving a record of what it removed: the model trained on C4 learned from a filtered world while the field's documentation described an unfiltered one.

The Data Provenance Initiative's licensing audit has the same reconstructive character at larger scale. Auditing nearly 2,000 open text datasets—more than 24,000 source documents in total across them—the initiative found that license information is most often absent entirely for crawled sources, that licenses are at times attached to collections but not to the individual sources whose terms they should govern, and that misdescription runs in both directions: sources described as more permissive than they are, and public-domain material unnecessarily restricted by dataset-level terms [4]. The audit's own framing is the supply-chain one—its authors describe the result as a transparency crisis in the "data supply chain" of AI—and its statistics give the upstream control question of Section 1.3 its empirical content: attribution fails at the majority of sources, verification fails wherever terms are misdescribed, and enforceability has nothing to attach to in either case [4].

It is worth being precise about why the audit came after the fact rather than before. The incentives of corpus construction point the wrong way: scraping is cheap, comprehensive, and fast, while permission tracking is slow, partial, and costly—and the training run that consumes the corpus can begin the day the crawl finishes. Auditing, by contrast, is publishable exactly when it is damning, so the institutions best placed to document their inputs have the least reason

to. This is the same asymmetry the validation literature documented in clinical AI, reproduced one stage earlier in the chain: the defect is not that documentation is technically hard, but that its costs fall on the party that benefits from its absence. Governance instruments—mandatory disclosure tiers, procurement requirements that condition acquisition on documentation—exist in mature supply chains precisely to override that incentive, and the upstream record shows what their absence has produced.

The current partial correction—the spread of opt-out signals and licensing deals—has the structure of a bargaining process conducted under the asymmetry the audits document. A content owner can exclude future crawls, but has no practical way to know whether past inclusion occurred, no inventory of what was taken, and therefore no basis for pricing what was lost; a model developer can offer a license for future use, which costs little against the alternative of litigation over the past. The NYT complaint crystallizes the endgame of that imbalance: when voluntary signals failed to produce accounting, the recourse is a demand, backed by court process, for precisely the inventory the supply chain never kept [1]. Read this way, the litigation is not an anomaly interrupting an otherwise functioning market; it is what markets do when the goods are traded without documentation and the audit arrives after consumption. The upstream lesson for governance is correspondingly blunt: documentary control points must be installed before the bargaining, not litigated after it, because every year of undocumented training increases the stock of knowledge whose provenance can never be established retroactively [4].

3. The Midstream: Four Injection Channels of Unequal Reversibility

3.1. Retrieval: The Fully Reversible Channel

Retrieval-augmented generation keeps knowledge outside the parameters entirely: documents live in an index, the model consults them at query time, and generation is grounded in what was retrieved [6]. Governance treats this channel as familiar territory. Correcting the index corrects the model's knowledge instantly and completely; removing a document removes its influence; licensing attaches to identifiable items. The channel's governance problems—index quality, retrieval faithfulness—are ordinary information-management problems. This is why staleness and copyright pressure in practice push deployments toward retrieval: it is the only channel where the traditional toolkit still works as designed.

The channel's origin paper is worth rereading as a governance document, because its design

choices are governance choices. Lewis et al. demonstrated RAG on knowledge-intensive tasks—open-domain question answering among them—showing that a generator conditioned on documents retrieved from a Wikipedia index reached competitive accuracy while the knowledge itself remained a separate, swappable component [6]. The architectural separation is what the governance analysis prizes: knowledge lives in an enumerated index whose contents, versions, and access terms can be stated; updates are index operations with predictable scope; and every generated claim has an identifiable evidentiary source that preceded it. None of this makes retrieval a complete control point—Section 3.1's boundary statement already limited the claim to evidence rather than use—but it supplies the precondition every other channel lacks: the object of governance exists as a list. When a regulator, court, or operator asks "what does the system know and where did it come from?", the retrieval channel is the only one of the four that can answer at all [6].

The channel's governance boundary deserves an honest statement, because retrievability is not faithfulness. Grounding generation in a retrieved source does not guarantee the source is represented accurately—the model can mischaracterize what it retrieved—and the index inherits whatever defects its corpus carries, including the upstream documentation failures of Section 2. What retrieval guarantees is narrower and still decisive: the evidence in the system is identifiable, correctable, and removable as a matter of record-keeping, even where the model's use of that evidence remains imperfect. In supply-chain terms, retrieval relocates the hard problem from unmanageable (distributed parameter change) to merely difficult (retrieval and generation quality)—and relocating a problem to where existing governance applies is exactly what supply-chain regulation has always done.

3.2. Model Editing: The Partially Reversible Channel

Model editing techniques attempt to retrofit reversibility onto parameters. ROME located the weight positions mediating a factual association and rewrote them with a rank-one update [7]; MEMIT scaled the approach to thousands of associations at once [8]. As governance instruments, however, the editing family carries a documented defect: edits do not propagate. On the MQuAKE benchmark, models whose underlying fact was corrected still answer multi-hop questions with the pre-edit answer, because the correction never travels along the chains of inference that the fact supports [9]. Knowledge-conflict studies complete the picture: when edited or retrieved content contradicts parametric knowledge, models behave inconsistently, sometimes deferring to stale parameters against explicit evidence [10]. Editing, in short, is reversible in the small and unreliable in the large—a channel where "correction" is locally

verified and globally unverified.

The methodological lineage explains why the verification gap took the field by surprise. ROME's contribution was diagnostic as much as corrective: using causal-mediation analysis—corrupting inputs and tracing which internal activations carry the factual recall—it located the specific mid-layer feed-forward positions where a subject–relation–object association is stored, and demonstrated that a rank-one weight update there could implant or replace a single fact [7]. The verification the method offers is rigorous but pointwise: after the edit, the target probe ("The Eiffel Tower is located in") returns the new answer, and neighboring associations are spot-checked for collateral damage. MEMIT generalized the operation, distributing thousands of edits across the layers the analysis identified, which is what moved editing from a laboratory curiosity toward an operation one could imagine running against a deployed model [8]. Both papers verified exactly what their framings suggested—local insertion and local side effects—and both are honest about the scope. The governance failure came later, when the operation's pointwise verification was quietly read as a system-level guarantee: a patch whose regression tests pass at the patch site is not thereby safe in a dependency graph, and facts are nothing but dependency graphs. MQuAKE [9] supplied the missing integration test, and the field learned what software engineering has known since the first regression: local unit tests do not certify global behavior [7-9].

The governance imagination that editing research invites is worth naming, because the field repeatedly reaches for it: the edit as patch, a correction issued against a deployed model the way a security fix is issued against software. MEMIT's scaling result is what makes the patch metaphor plausible—thousands of updates, issued at once [8]. What MQuAKE's result is what makes it premature: a patch that does not propagate through the dependencies of what it changes is not a patch in any sense software engineering would recognize, and knowledge is almost nothing but dependencies—a corrected founding date implies corrected ages, durations, and sequences that no current editing method traces [9]. Until edit operations carry their dependency structure, the patch metaphor should be treated as aspiration rather than capability, and corrections that matter legally should be routed through channels whose propagation is guaranteed—that is, back through the index or the corpus.

3.3. Pre-training and Fine-tuning: The Effectively Irreversible Channel

What pre-training injects, nothing yet reliably extracts. Carlini et al. demonstrated that language models memorize training data deeply enough to regurgitate it—verbatim, including personally identifiable records—under black-box queries [11], and Brown et al. sharpened the

governance meaning of that fact: memorization is not itself the harm, but memorized content whose collection context promised privacy constitutes a standing disclosure waiting for the right query [12]. The channel's reversibility is effectively nil at pre-training scale; fine-tuning sits between channels, injectable knowledge that unlearning research targets but, as Section 5 shows, cannot yet verifiably remove. Fine-tuning's governance position is also the least examined: it is the channel through which organizations install their private, often regulated, knowledge into models, yet it carries none of the documentation tradition that upstream datasets have begun to acquire—an enterprise fine-tune is frequently the least documented acquisition of knowledge in the entire chain. Table 2 grades the channels.

The memorization evidence also quantifies the channel's exposure in a way that matters for governance planning. Carlini et al.'s extraction methodology—queries designed to elicit high-perplexity continuations, which surface memorized training text against the model's fluency prior—recovered verbatim personally identifiable records from a production-scale model, and the study's two structural findings are the ones a data controller must plan around: memorization per example increases with model scale, so the exposure grows with every capability generation; and repetition of an example in training drives memorization steeply, which makes duplication of sensitive records a controllable risk factor—the rare upstream decision that measurably changes the downstream harm [11]. Brown et al. then supplied the definitional correction that turns these results into a governance standard rather than a curiosity: privacy, they argue, cannot be evaluated in the abstract as "is this string in the weights," but only relative to the context in which the data was collected—a public court record and a private medical record may both sit in the weights, but only one was collected under a promise of confidentiality, and only the latter constitutes disclosure when extractable [12]. The standard relocates the compliance question from corpus statistics to collection context—which is to say, once more, to upstream documentation. A controller who cannot state under what expectations each corpus was gathered cannot say whether its model's memorization constitutes a standing breach; the midstream defect is inheritable from the upstream one [11,12].

Table 2. The four knowledge-injection channels graded by governance properties.

Channel	Knowledge location	Reversibility	Verified correction	Recall instrument
Retrieval [6]	External index	Full	Immediate, complete	Drop document
Model editing [7,8]	Weights (localized)	Partial	Local only; ripples fail [9]	Un-edit (unproven)
Fine-tuning	Weights (distributed)	Low	Unlearning, fragile (Section 5)	Unlearning
Pre-training [11]	Weights (entangled)	Effectively none	None demonstrated	Retrain (prohibitive)

The grading exposes the midstream's governance inversion: the easiest channel to inject at scale—pre-training—is the hardest to govern, while the governable channel—retrieval—is the one treated as an auxiliary. Deployment architecture is quietly a governance decision, and it is rarely framed as one.

The inversion has a specific casualty worth naming: the fine-tuning row of Table 2. Fine-tuning is the channel through which enterprises install their own regulated knowledge—customer records, internal policies, proprietary research, in many cases personal data subject to the obligations of Section 5—into a model, and it inherits pre-training's distributed irreversibility in miniature: the knowledge is in the weights, entangled with the base model's, removable only by unlearning of the fragile kind Section 5 measures. Yet the governance discussion concentrates on pre-training corpora and on retrieval indexes, while the enterprise fine-tune—the one injection most directly under a single controller's authority, and therefore the one where documentary discipline is actually achievable—proceeds with no datasheet, no channel statement, and no deletion story. A controller who cannot govern anything else can still govern what it fine-tunes on; that this channel is the least documented in practice is not a technical constraint but a choice the governance instruments have not yet forced anyone to revisit.

4. The Maintenance Problem: Staleness, Conflict, Hallucination

Injected knowledge ages, and the aging is measurable. TemporalWiki constructs a lifelong benchmark from consecutive Wikipedia snapshots and shows that models trained at one moment systematically fail on facts established or changed after it, with continued pre-training required—precisely the least reversible channel—to keep pace [13]. Knowledge conflicts are the second maintenance failure: Section 3.2's evidence that models defer inconsistently between fresh and parametric knowledge [10] means staleness is not merely missing information but actively misleading information the model may prefer. Hallucination is the downstream

symptom common to both: Ji et al.'s survey organizes the phenomenon—generation unfaithful to sources, whether contradicting them or fabricated from nothing—along with causes and mitigations, and its taxonomy has become the field's shared language for the defect [3].

TemporalWiki's measurement protocol is the part with the sharpest governance implication, because it shows the decay is not merely real but invisible to the affected party. The benchmark trains two models on adjacent Wikipedia snapshots and evaluates each on the other's snapshot, in both temporal directions, so that the measured quantity is precisely the knowledge difference between two moments of a changing world [13]. The finding that stings: a model degrades measurably on facts that changed after its training cutoff, yet nothing in its behavior flags the degradation—the model answers fluently and confidently from frozen premises, and only an external benchmark keyed to the world's own versioning reveals that the answers have gone stale [13]. Knowledge in parameters, in other words, decays on a schedule the operator cannot observe without constructing a diff against the world's current state; knowledge in an index decays exactly as fast as its sources do, and the diff is the index update itself. That asymmetry—identical decay, opposite observability—is arguably the strongest architectural argument for the retrieval channel in the entire governance literature, and it is invisible in benchmarks that evaluate capability at training time rather than knowledge maintenance over time [13].

The three maintenance failures are usually catalogued side by side, but the supply chain reading gives them a causal order that cataloguing obscures. Staleness originates upstream and midstream: the world changed and the injected knowledge did not, because the injection channel froze its contents at crawl or training time. Conflict arises next, in the gap between the frozen parameters and anything fresher the system consults—the model must adjudicate between its own aging knowledge and newer evidence, with the adjudication behavior documented as inconsistent. Hallucination compounds both: a model asked about a world its parameters no longer describe must either abstain—performance that benchmarks show is unreliable—or interpolate, and interpolation over stale premises is precisely where unfaithful generation concentrates. The ordering matters for governance because it locates the intervention points: staleness is addressed at injection architecture, conflict at adjudication behavior, and hallucination only at output—the last and weakest place to intervene. The maintenance record supports one further governance conclusion: knowledge inside parameters degrades on its own schedule, unobservable without dedicated benchmarks, while knowledge in retrieval degrades exactly as fast as its sources—and visibly. Whatever else the channels differ in, they differ in whether decay is auditable.

5. The Recall Problem: Removal Without a Mechanism

Recall is where the supply chain currently fails hardest, because the demand for it is legal while the capability is technical. TOFU, the first unlearning benchmark with knowable ground truth—fictitious author profiles that the model was trained on and can be probed against—reports that existing unlearning techniques preserve model utility while largely failing to remove the targeted knowledge, and that the failure worsens as the volume to forget grows [14]. The privacy stakes inherit from the memorization evidence of Section 3.3: content that cannot be verifiably removed remains extractable [11,12], whatever the data controller's policy says.

TOFU's design is what makes its negative result evidentially strong rather than merely discouraging. By training models on 200 fictitious authors—roughly 4,000 question–answer pairs generated for that purpose—the benchmark manufactures the condition real deployments lack: the ground truth of what the model knows and therefore of what "forgotten" would mean [14]. Its evaluation plane is correspondingly differentiated: beyond target knowledge, it probes forget quality (including membership-inference probes that ask whether the forgotten data remains detectable in the model's behavior), model utility on general and related author performance, and the trade-off frontier between the two. The reported pattern is the unlearning dilemma in miniature: methods that preserve utility largely fail to remove (the unlearned knowledge remains recoverable or inferable), and methods that remove more thoroughly pay for it in collateral degradation of unrelated capability—with the dilemma worsening as forget volume grows [14]. For governance, TOFU supplies two things at once: the demonstration that current unlearning is not deployment-grade, and the evaluation protocol that any future claim of verified deletion would have to survive—probing, inference attacks, utility accounting [14]. A deletion standard, in other words, already has a candidate form; what the field lacks is models that pass it.

The statutory side has more structure than the literature's references to "the right to be forgotten" suggest, and the structure matters for engineering. The GDPR's erasure right (Article 17) operates alongside the storage-limitation principle (Article 5), which requires that personal data not be kept longer than necessary—two distinct obligations, of which the second binds continuously, not only upon request—and the enforcement regime reaches global turnover percentages, which is what gives the obligations their pricing power [15]. The contested question for large models is threshold: whether personal data embedded in weights constitutes "processing" of that data at all, an interpretation on which the erasure obligation's applicability to trained models turns, and on which regulators and scholars have not fully converged [15].

But the uncertainty cuts against the field rather than for it: a controller who cannot verify what personal content sits in the weights—the memorization problem of Section 3.3—cannot establish that none is there, and the practical compliance posture therefore defaults to treating parameters as a plausible processing surface with no working deletion mechanism. That is the recall gap in statutory terms: two binding obligations, one contested threshold, and zero verified mechanisms on the technical side [14,15].

Law does not wait for the engineering. The GDPR's right to erasure obliges controllers to delete personal data, with narrow exemptions, and a corpus is personal data in the statute's terms long before it is weights [15]; the NYT complaint's requests go further, seeking to force unwinding of what training did [1]. The mismatch is stark: the legal system already issues recall orders against a supply chain whose recall stage has no verified mechanism. Corpus-side withdrawal—removing a source before training—remains the only recall that works as intended, and it acts only on models not yet built.

For models already deployed, the honest summary of the field's evidence is that recall is currently performed in three ways that work and one that is demanded but not yet delivered. The ways that work: drop the retrieval index, which removes the knowledge from the system's evidentiary base instantly; retrain from a cleaned corpus, which works but costs as much as building the model; and do not train on the content in the first place, the preventive form that opt-out signals, licensing negotiations, and provenance audits all serve. The way that is demanded but not delivered is surgical unlearning—removing one source's influence from finished weights—which is exactly the capability TOFU measures and finds wanting [14]. The pattern across the four is instructive: recall capability is inversely proportional to how deep the knowledge sits, which is Table 2's reversibility grade read backward. Governance that takes the evidence seriously will therefore concentrate on the two ends of that gradient—making the preventive end cheap enough to be used, and demanding verification at the surgical end rather than accepting its promises.

6. Governance Across the Chain: Control Points

The governance instruments that exist map naturally onto chain stages, which is both their strength and their limit. Upstream, datasheets for datasets propose the documentary record the chain lacks: motivation, composition, collection process, and recommended uses attached to every dataset, in the manner of component datasheets in electronics [16]—the instrument that, had it existed at scale, would have answered the audit questions that Dodge [5] and the Data Provenance Initiative [4] had to reconstruct after the fact. Downstream, model cards give

released models a documented boundary of intended use and evaluated performance [17]. The EU AI Act overlays both ends with enforceable obligations: training-data summaries and record-keeping for general-purpose model providers, and logging duties downstream for high-risk deployers [18].

The documentation instruments are also worth reading at the granularity of their questions, because their questions are the supply chain's questions. Datasheets specify a dataset's motivation, composition, collection process, preprocessing, and recommended uses—each heading a field the C4 audit found missing in practice: the collection process question is the provenance record the corpora lack; the preprocessing question is the record of what filtering removed; the composition question is what would have exposed the register skew before training rather than after [16]. Model cards, symmetrically, attach to a released model its intended use contexts, evaluation data, and performance broken down by demographic and task group—a reporting structure whose disaggregation discipline is what makes decline-on-subpopulations visible instead of averaged away [17]. The two instruments bracket the chain at its documentable ends, which is exactly their limit: between the dataset that enters and the model that ships sit the injection channels, and no instrument in current standard use requires stating which channel carries which knowledge or how reversible that carriage is. The documentation tradition, in other words, has already solved the hardest problem—making disclosure habitual—and its unfinished work is extending the disclosure across the midstream [16,17].

Two gaps remain that documentation cannot close. First, the midstream lacks its control point entirely: no standard requires a deployment to state which channel carries which knowledge class, though Table 2's reversibility grading is exactly the kind of statement that could be required—a system carrying personal or licensed knowledge in the least reversible channel should have to say so. The AI Act's training-data summary obligation for general-purpose providers is the nearest existing instrument, and its formulation—a public "sufficiently detailed summary" of training content—will be tested against precisely this question: whether a summary that omits provenance quality and channel reversibility can be considered sufficiently detailed at all [18]. Second, recall lacks verification: deletion claims are currently unfalsifiable without probing protocols of the kind TOFU introduced [14], and a governance regime that accepts unaudited deletion claims replicates upstream's original sin—accepting inputs on faith. The supply chain reading turns these from philosophical worries into specification gaps with natural owners: standards bodies for the documentation tiers, model providers for reversibility

disclosure, deployers for verified deletion.

The AI Act's obligations, read closely, already sketch the midstream disclosure this survey argues for, though in entry-event rather than channel terms. The Act's documentation duties for general-purpose model providers include technical documentation covering the training stage and a publicly posted "sufficiently detailed summary" of training content—the first time European law has required model developers to describe what went into the weights at all [18]. Its record-keeping obligations (automatically generated logs for high-risk systems) and post-market monitoring duties extend the governance horizon from authorization to lifecycle, which is the same reorientation the recall problem demands: a model whose knowledge can be corrected or removed is governed by a different regime than one whose knowledge is frozen at release. What the Act does not yet do—the gap this survey's Table 2 identifies—is require the channel-level statement: which knowledge classes entered through which injection channel, and what each channel's reversibility grade implies for the operator's correction and deletion obligations. That omission is where the specification work remains: the Act created the disclosure habit; the supply chain framework specifies what the disclosure must say to govern knowledge rather than merely document it [18].

7. Conclusion

This survey organized the knowledge governance of large language models as a supply chain and reached three findings. First, the upstream is structurally unauditable today: the field's own audits document corpora with no provenance discipline [5] and datasets that mostly fail to convey their licensing terms [4]. Second, the midstream's four injection channels are graded—retrieval fully reversible [6], editing locally reliable but globally leaky [7-9], pre-training effectively irreversible [11]—and deployment practice inverts the governance order, concentrating knowledge where it is hardest to govern. Third, recall is demanded by law but not delivered by technique: unlearning remains fragile at scale [14], erasure obligations bind regardless [15], and litigation has arrived to test the gap [1]. The control points that close these gaps—provenance records [16], reversibility disclosure, verified deletion—are documentary and procedural rather than heroic, which is precisely why they are achievable.

The sequencing of that work is not arbitrary, because the control points depend on each other in one direction. Verified deletion is the glamorous problem, but it is downstream of reversibility disclosure—an operator cannot be held to delete what no record says was injected—and reversibility disclosure is downstream of provenance documentation, since "which knowledge entered" presupposes "what was known about what entered." The

dependency chain mirrors the stages themselves, which is the deepest sense in which the supply chain is the right frame: upstream defects do not stay upstream, they propagate into every later control's preconditions, exactly as a defective input lot propagates through a manufacturing process. Governance programs that begin at recall—demanding deletion first—therefore invest at the point of maximum dependency and minimum feasibility, while programs that begin at documentation make every later stage cheaper and more verifiable. The recommendation this survey's evidence supports is accordingly ordered: make the documentary habit universal at the upstream [16,17], extend it across the midstream as channel-level reversibility disclosure [18], and let the demand for verified deletion [14,15] concentrate the field's research investment where the engineering is genuinely missing.

The framework travels to any technology whose output is learned rather than recorded—recommender systems, foundation-model derivatives, embodied controllers. For language models specifically, the highest-value open problems follow the chain: upstream, licensing infrastructure that makes permission machine-checkable; midstream, ripple-propagating edit methods that would move model editing from Table 2's "partial" grade toward the retrievable column; downstream, maintenance benchmarks that treat staleness as a reportable property; and at recall, deletion that survives adversarial probing. The NYT lawsuit asked a court to do what the technology cannot; building the technology that can is the field's share of the answer.

References

[1] The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. 2023).

[2] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[3] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, Y., Xu, D., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[4] Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2023). The Data Provenance Initiative: A large scale audit of dataset licensing & attribution in AI. arXiv preprint arXiv:2310.16787.

[5] Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information

Processing Systems 33.
[7] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. In Advances in Neural Information Processing Systems 35.
[8] Meng, K., Sen Sharma, A., Andonian, A., Belinkov, Y., & Bau, D. (2023). Mass-editing memory in a transformer. In The Eleventh International Conference on Learning Representations (ICLR).
[9] Zhong, Z., Wu, Z., Manning, C. D., Potts, C., & Chen, D. (2023). MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.
[10] Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In The Twelfth International Conference on Learning Representations (ICLR).
[11] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633-2650).
[12] Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022). What does it mean for a language model to preserve privacy? In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 2280-2292).
[13] Jang, J., Ye, S., Lee, C., Lee, S., Yoon, D., Seo, M., & Yang, S. (2022). TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing.
[14] Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., & Kolter, J. Z. (2024). TOFU: A task of fictitious unlearning for LLMs. In First Conference on Language Modeling (COLM).
[15] European Parliament and Council. (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). Official Journal of the European Union, L 119.
[16] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.
[17] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*) (pp. 220-229).
[18] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

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

From Benchmark to Bedside: Mapping the Translational Gaps of Artificial Intelligence in Clinical Medicine

Bohan Li,Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: August 23 2026

Revised: August 30, 2026

Accepted: August 30, 2026

Published online: August 30, 2026

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

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

Abstract. Artificial intelligence (AI) has reached specialist-level performance across a widening range of clinical tasks, yet celebrated benchmark results have repeatedly failed to translate into patient benefit. This survey reviews 29 core sources (2016–2024) through an original framework of three translational gaps—validation, utility, and governance—that jointly explain the distance between benchmark scores and bedside value; reported figures are quoted only from peer-reviewed studies, pivotal trials, or first-party regulatory documents. Capability measurement has matured across three generations—specialist image classifiers, examination-scale language models exceeding 86% on USMLE-style questions, and workflow-level deployments such as autonomous diabetic-retinopathy screening and ambient clinical documentation. Each gap independently severs the link between performance and benefit: deployed models degrade under dataset shift, an externally audited sepsis model fell far below its reported accuracy, nearly half of the citations in generated medical content can be fabricated, and human–AI collaboration can underperform either party alone. Governance responses—risk-tiered regulation, equity audits, and post-market surveillance—are converging on medicine's established logic: grading AI claims with the same evidence hierarchy that evidence-based medicine applies to therapies.

Keywords: *medical imaging; large language models; external validation; algorithmic bias; regulatory approval*

1. Introduction

1.1. Background and the Central Contradiction

Medicine has become the most visible testing ground for artificial intelligence (AI). Deep learning systems now read retinal photographs, chest radiographs, histopathology slides, and dermatological images; large language models (LLMs)—neural networks trained on web-scale text—answer licensing-examination questions, draft clinical documentation, and summarize patient records. Commentators have described this convergence as the arrival of "high-performance medicine," in which machine perception and recall augment clinical judgment across specialties [1]. Health systems have acted on the promise: AI-enabled products are embedded in radiology workflows, emergency departments, and primary-care clinics. The speed of adoption is unusual even by the standards of recent medical technology; the deeper novelty is that medicine, of all professions, already possesses a mature institutional answer to the question of what counts as evidence—and AI systems are arriving faster than that machinery has learned to grade them.

That institutional answer is worth describing at the outset, because this survey's entire argument is that AI evaluation is converging on it from first principles. Evidence-based medicine ranks forms of evidence by their resistance to self-deception: mechanistic reasoning at the bottom, retrospective observation above it, controlled comparison higher still, and pre-registered randomized trials with independent surveillance at the top—paired with the institutions that operate the hierarchy, from ethics boards to trial registries to pharmacovigilance systems [1]. The hierarchy was not designed for software, but its logic is software-agnostic: it asks of any claim not how impressive it is but what would have revealed it to be wrong. A benchmark score is impressive in exactly the way a mechanistic argument is impressive—and, as the sections that follow document, it fails in exactly the way mechanistic arguments fail: persuasively, retrospectively, and without any procedure that would have caught its errors before patients absorbed them. The survey's governance conclusion can therefore be stated here and defended for the rest of the paper: the question is not whether medicine needs new institutions for AI, but whether AI evaluation will adopt the institutions medicine already built, or repeat the centuries it took to construct them.

The promise, however, has coexisted with a persistent pattern of disappointment, and the pattern is not accidental. In 2017, a Nature study reported that a deep neural network classified skin cancer at dermatologist-level performance across 21 disease categories, an achievement widely covered as a turning point for the field [2]. In the same year, an investigative report

documented that IBM's Watson for Oncology—a system deployed to recommend cancer treatment—had generated "unsafe and incorrect" therapy suggestions in hospital settings, contradicted by the physicians meant to use it [3]. The juxtaposition frames this survey's central question: why do headline performances on clinical benchmarks so often fail to become safe, useful, and equitable bedside systems?

This survey argues that the failure is structural rather than incidental, and that it decomposes into three independent translational gaps: the validation gap (benchmark performance does not survive contact with prospective clinical data), the utility gap (technical accuracy is not patient benefit), and the governance gap (a validated system is not automatically an accountable or equitable one). Each gap has its own failure evidence, its own corrective tradition, and its own open problems; conflating them—which much of the public debate does—produces both exaggerated hope and indiscriminate skepticism.

The independence of the gaps is the framework's operative claim, and it cuts both ways. Because they are independent, no amount of progress on one closes another: a model validated prospectively at multiple sites has addressed the first gap and none of the others; an interventional outcome study of an unvalidated model measures a moving object; a regulated deployment of a utility-blind system enrolls regulators in harms it does not prevent. Each of the field's celebrated failures can be located on exactly one gap, which is why single-lever explanations of the benchmark-to-bedside distance keep failing—the deployment pessimist points at validation evidence, the adoption optimist at utility trials, and each dismisses the other's gap as an implementation detail. The gaps also differ in who can close them: the validation gap is closed by evaluators and regulators, the utility gap by trialists and health systems, the governance gap by legislators and professional bodies—an allocation of labor the literature rarely makes explicit but that determines which failures should be assigned to which institutions. The three gaps are, in short, the field's problem structure, and the survey organizes its evidence to show that they are real, measured, and separately actionable.

1.2. Survey Methodology

We conducted a targeted literature search across PubMed-indexed general and specialty journals (Nature Medicine, JAMA, NEJM, The Lancet Digital Health, npj Digital Medicine, PLOS Medicine), AI conference proceedings, and first-party regulatory publications (U.S. FDA, European Union, WHO). The search covered 2016–2024 for technical work, 2016 being the year the modern clinical-deep-learning literature opened with large-scale retinal screening. Candidates were retained if they (i) reported empirical clinical or examination performance of

AI systems, (ii) provided independent or adversarial evaluation of deployed systems, or (iii) constituted primary governance instruments (regulations, guidance, official device lists). Purely speculative commentary was excluded. Reported figures are quoted only from peer-reviewed publications, pivotal trials, or first-party documents—never from secondary commentary. This procedure yielded the 29 core sources organized below.

The discipline behind that last sentence bears explicit description, because surveys of medical AI are unusually exposed to citation drift. Every quantitative claim in this survey was traced to the document that produced it: performance figures to the original benchmark or trial publications, audit findings to the auditing studies, regulatory counts to the regulator's own published list, and legislative obligations to the legislative texts themselves. Author lists and bibliographic details of every source were verified against its primary record before inclusion, and sources that could not be verified with confidence were excluded rather than cited from memory. Where this survey draws an inference the source does not state—that a finding constitutes a validation failure, for instance—the inference is presented as interpretation, not attributed to the source. The survey holds itself to the evidentiary standard it argues the field should hold AI systems to: traceable claims, primary provenance, and no number whose origin cannot be named in one step. Readers who find a figure here can follow it to its document; that is the property Sections 3 and 4 will argue deployed clinical AI currently lacks for its own claims.

1.3. Positioning Against Existing Surveys

The field's self-understanding is framed by several recent reviews. Rajpurkar et al. survey AI in health and medicine by modality—image, text, and multimodal systems—and catalogue technical progress and deployment barriers [4]. Thirunavukarasu et al. review LLMs in medicine specifically, organizing the material by clinical task [5]. Kelly et al. articulate the challenges of delivering clinical impact—data quality, workflow integration, and regulation—providing the closest prior framing, though written before the LLM era and without the quantitative failure evidence that has since accumulated [6]. All three organize the field by what AI is (modalities, tasks) or what obstacles exist (challenge lists); none treats the inferential status of performance claims as the organizing problem. Table 1 contrasts their coverage with this survey's, whose distinctive commitment is to grade every performance claim by its translational status—the logic evidence-based medicine (EBM) already applies to therapies.

Table 1. Coverage comparison between recent surveys of AI in medicine and this work.

Dimension	Rajpurkar et al. [4]	Thirunavukarasu et al. [5]	Kelly et al. [6]	This survey
Primary scope	Multimodal techniques	LLMs by clinical task	Barriers to impact	Benchmark-to-bedside evidence
Organizing logic	Modality	Task category	Challenge list	Three translational gaps
Failure evidence	Selected examples	Brief	Conceptual	Quantified: shift, hallucination, external audit
Governance treatment	Brief	Brief	Conceptual	Regulation, equity, post-market surveillance

The remainder follows the gap structure. Section 2 reviews what has actually been measured, across three generations of capability. Section 3 presents the three gaps and the evidence for each. Section 4 analyzes the governance responses closing them, and Section 5 concludes.

2. Measured Capability: Three Generations of Clinical AI

2.1. The Specialist Era: Deep Learning on Medical Images

The modern capability story begins with specialist image classifiers. Gulshan et al. trained a convolutional network on more than 128,000 retinal fundus photographs and detected referable diabetic retinopathy with an area under the receiver operating characteristic curve (AUC) above 0.99 on two independent validation sets—performance matching ophthalmologists [7]. Esteva et al. classified skin cancer across 21 categories at dermatologist level using 129,450 clinical images [2]. De Fauw et al. translated optical coherence tomography scans into referral recommendations matching expert clinicians on sight-threatening retinal disease, and showed the representation transferred across scanning devices [8]. Chest radiography followed: CheXNet detected pneumonia at radiologist level [9], and the CheXpert dataset—224,316 radiographs of 65,240 patients with automated uncertainty-aware labels—industrialized benchmark construction for the modality [10].

Three properties defined the era. Measurement was single-task: one disease, one image type, one decision. Ground truth was retrospective: labels extracted from existing reports rather than from what happened to patients. And the comparator was the specialist: success meant matching clinicians on their home turf, not improving outcomes. These properties made spectacular headline results possible—and, as Section 3 shows, planted the first translation gap. They also set a template that survives into the present: the field's most-cited capability claims are still, at bottom, agreement statistics computed on archives, however sophisticated the architecture

reading the archive has since become.

The archive structure of the era's landmark studies rewards a second look, because it determined what the studies could and could not show. Gulshan et al.'s retinal classifier was validated on two data sets—one drawn from the same screening network as training data, one from a different network—both curated and retrospective: images already captured, already graded, chosen to measure agreement rather than consequence [7]. Esteva et al.'s dermatology classifier was tested on photographs already verified by biopsy or follow-up, a design that guarantees a clean reference standard and guarantees equally that the test distribution is nothing like a clinic's intake, where prevalence is lower, quality is uncontrolled, and the difficult cases were already referred to specialists [2]. De Fauw et al.'s OCT referral system went furthest of the three toward clinical realism—its cohort was drawn from a real care pathway, and the analysis included simulated deployment scenarios—but its comparison target remained the clinicians' own decisions rather than downstream outcomes [8]. None of these qualifications diminishes the studies; they were measurement milestones executed with unusual care. The point is what their designs share: in every case the patient had already been triaged, imaged, and diagnosed by the ordinary machinery of care before the AI entered the pipeline, and the AI's agreement with that machinery was the entire result. What happens when the AI enters before the machinery—not after it—is the subject of Section 3.

2.2. The Generalist Era: Medical Question Answering

The second generation shifted from images to language. MedQA assembled large-scale question answering from professional medical examinations used in the United States and China, operationalizing clinical knowledge rather than perception [11]. On the MultiMedQA suite built from six such datasets, the instruction-tuned Flan-PaLM reached 67.6% on MedQA—surpassing prior state of the art by more than 17 points—and its instruction-tuned derivative Med-PaLM was the first model to exceed the USMLE passing threshold [12]. Its successor Med-PaLM 2 reached 86.5%, and in blinded comparison physicians preferred its consumer-question answers to those written by physicians themselves on eight of nine evaluated axes [13]. GPT-4, with no medical specialization at all, exceeded the USMLE passing score by over 20 points and was markedly better calibrated than GPT-3.5 [14]; ChatGPT itself performed at or near the passing threshold across all three USMLE steps [15].

Yet the same studies that established these scores qualified them. Human evaluation of long-form answers found Flan-PaLM's outputs repeatedly inferior to clinicians' on factuality and potential harm even where multiple-choice accuracy was state of the art [12]. The measurement

object had quietly migrated: from whether a model answers like a specialist on a curated question, to whether it behaves like a clinician on an open one—and the second claim was not established by the first.

The human-evaluation arm of the Flan-PaLM and Med-PaLM 2 studies is the era's most instructive evidence precisely because it measured the migration directly. Alongside the multiple-choice results, the investigators had clinicians rate long-form answers on nine axes—factuality, potential for harm, comprehension, and related quality dimensions—and the pattern held across both model generations: answers that were state of the art on examination scoring remained repeatedly rated inferior to clinicians' own on the axes that matter to a patient [12]. Med-PaLM 2 closed much of the gap—in blinded comparison its answers were preferred to physician-written ones on eight of nine axes—but the residual axis tells the honest story: the axis on which the model still trailed was potential for harm, the one dimension where an error is not an inconvenience but an injury [13]. An examination cannot register that dimension; a multiple-choice key has no row for it. The generalist era's deepest result is therefore not the 86.5% but the demonstration that examination scores and clinical-answer quality are different measurements that happen to correlate loosely—and that a field reading the first number as evidence about the second was making a categorical error the studies themselves had already refuted in their own appendix tables [12,13]. The era also internationalized evaluation: MedQA drew its questions from professional examinations on both sides of the Pacific, and the cross-country hallucination benchmarks discussed further in Section 3.2 ask whether models trained on predominantly English, Western corpora remain sound when questioned under other jurisdictions' testing standards [11]. Capability, in short, is now measured globally—but mostly with the Anglophone model as protagonist.

2.3. The Workflow Era: From Models to Deployed Systems

The third generation measures systems, not models. IDx-DR became the first FDA-authorized fully autonomous diagnostic AI, referring or clearing diabetic-retinopathy patients in primary-care offices without any physician interpretation, on the strength of a prospective pivotal trial [16]. Ambient AI scribes—LLMs that listen to clinical encounters and draft documentation—represent the fastest LLM deployment to date: one deployment study across a large integrated care system reported 968 physicians with at least 100 encounters each and an estimated 15,000 hours of documentation time saved over 2.5 million uses [17].

The two workflow deployments are also, read carefully, a controlled lesson in how to make translation succeed—because they chose the two easiest cells of the task grid and still illustrate

every remaining gap. Autonomous screening succeeded at IDx-DR only after a pivotal trial designed to regulatory standard, a task with a binary refer/clear decision, a modality with a stable imaging protocol, and a care setting where the consequence of error is another appointment rather than a missed diagnosis [16]. Ambient scribes succeeded first because their output is draft: a clinician reviews and signs every note, so the human-in-the-loop is structurally built in and errors are absorbed by the professional rather than transmitted to the patient [17]. Neither deployment required solving the hard problems of Section 3—they were selected, deliberately or not, so that validation is well-defined and error is recoverable. The lesson generalizes in an uncomfortable direction: the workflow tasks now measured are exactly the ones where the translational gaps are narrowest, while the tasks where AI could most change outcomes—autonomous management of undifferentiated patients, high-stakes treatment selection—remain outside both the benchmark literature and the deployment literature. The workflow era measures where translation is easy; patients benefit where translation is hard [16,17]. Table 2 summarizes the three generations. Figure 1 situates them within the translational-gap framework of this survey.

Table 2. Three generations of clinical AI capability measurement.

Generation	Representative systems	Measurement object	End point
Specialist vision (2016–2019)	Retinal DR [7], skin cancer [2], OCT referral [8], chest X-ray [9,10]	Single-task image classifier	Agreement with specialists (AUC)
Generalist QA (2021–2023)	MedQA [11], Med-PaLM [12], Med-PaLM 2 [13], GPT-4 [14]	LLM on examination questions	Accuracy; physician preference
Workflow systems (2018–)	Autonomous screening [16], ambient scribes [17]	Deployed system in clinical process	Task completion; time saved

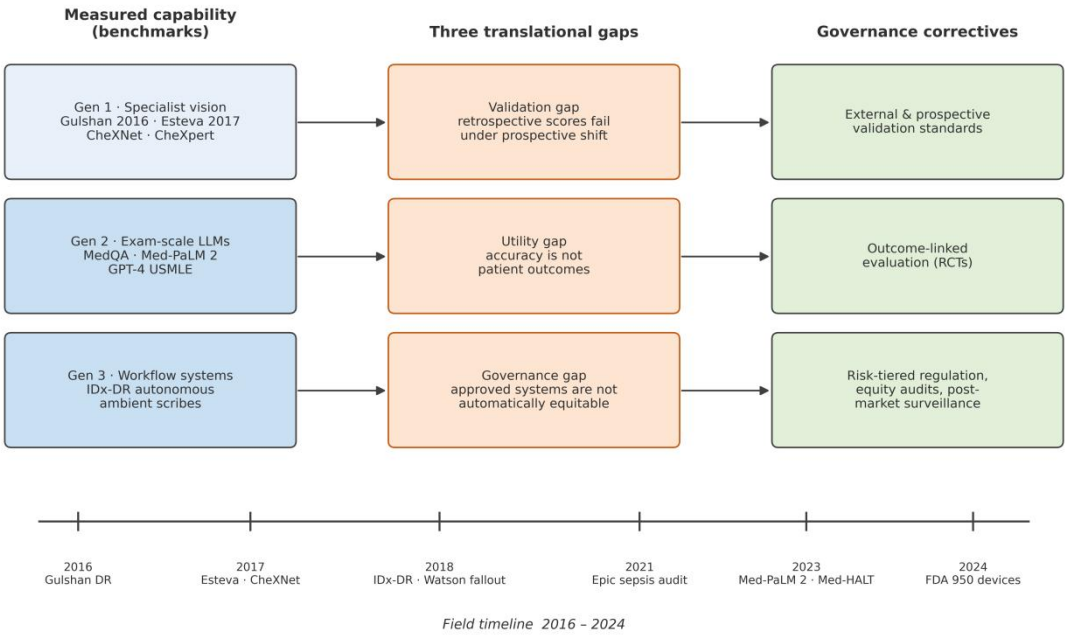

Figure 1. The three translational gaps of clinical AI. Left: three generations of measured capability. Middle: the independent gaps separating benchmark performance from bedside value. Right: the governance correctives each gap demands. Bottom: the field's timeline, 2016–2024.

2.4. The Capability Boundary

Aggregate capability, however, is bounded in ways aggregate scores conceal. Examination performance measures closed-book knowledge under standardized prompts—the setting examinations use precisely because it is not practice [12,14]. Long-form evaluation repeatedly exposed gaps in reasoning, currency of knowledge, and calibrated uncertainty that multiple-choice accuracy does not register [12,14]. Whole regions of practice remain thinly measured at all: dosage and drug-interaction reasoning under real-world polypharmacy, rare-disease presentation, multimodal integration of imaging with narrative records, and the epistemics of not knowing—a clinical skill examinations reward only in passing. And deployment evidence remains concentrated where measurement is easiest: documentation is now measured at industrial scale [17], while diagnostic autonomy beyond retinal screening remains rare [16]. There is an asymmetry worth naming: the tasks AI has penetrated fastest are those whose outputs are most easily reviewed—text a clinician edits, images a radiologist confirms—while the tasks patients most need help with are precisely those whose outputs are hardest to check. What AI systems can do, in short, has been measured well; what deployed systems do to patients has not. That distinction is the subject of the next section.

The boundary analysis also predicts where the field's measurement effort will concentrate next, and why that prediction is itself uncomfortable. The pattern across all three generations is

that capability follows measurability: imaging benchmarks preceded imaging deployment, examination benchmarks preceded documentation deployment, and the tasks currently unmeasured—polypharmacy reasoning, undifferentiated presentations, epistemic humility about not knowing—are precisely the tasks without a data collection tradition [12,14]. If the pattern holds, the next capability frontier will be set by whatever new measurement infrastructures are built rather than by modeling breakthroughs, which inverts the field's usual narrative of algorithm-driven progress. It also creates a quiet selection pressure worth naming: tasks that resist measurement will continue to resist deployment, not because they are harder for AI but because nobody can claim anything about them—and the aggregate effect is a clinical AI portfolio optimized for the measurable rather than the important. Recognizing that pressure is the boundary section's practical yield: what the field chooses to build benchmarks for is, increasingly, a public-health decision, and it is currently being made by default rather than by design [12,14].

3. The Three Translational Gaps

3.1. The Validation Gap: Retrospective Scores Under Prospective Conditions

The first gap is metrological. A retrospective benchmark measures performance on a static, curated dataset; a deployed model meets a moving clinical world. Zech et al. provided the canonical demonstration: chest-radiograph classifiers exploited hospital-specific signatures—platform artifacts and disease-prevalence patterns of their home institutions—and performance consistently degraded when tested at other hospitals, because the model had partially learned where an image came from rather than what it showed [18]. The critique extends to deployed systems at scale. An independent external audit of Epic's widely implemented sepsis-prediction model—deployed across hundreds of U.S. hospitals with a reported AUC of 0.76–0.83—measured an AUC of 0.63, and found the model missed two-thirds of sepsis cases while alerting on many patients who never developed it [19]. Dataset shift itself has been given a clinical taxonomy—changes in practice, population, measurement, and label conventions—along with the argument that safe deployment requires informed clinician-users plus technical governance, because shift cannot be eliminated, only monitored [20].

In EBM terms, a retrospective benchmark carries roughly the weight of mechanistic plausibility: necessary, encouraging, and not licensure-grade. The corrective is external and prospective validation on moved data—a standard the field has named far more often than it

has funded. The persistence of the gap has an incentive explanation: public benchmarks are cheap, fast, and competitive, and they reward the institutions that build them; external validation is slow, expensive, unglamorous, and its findings—whoever pays for them—are publishable precisely when they are damaging. No amount of methodological exhortation overcomes that asymmetry; only funding structures and regulatory requirements can, which is why the validation gap is ultimately a governance problem even though its failure mode is statistical.

Zech et al.'s demonstration repays mechanistic unpacking, because it shows the failure is not poor engineering but rational learning of spurious structure. Their classifiers, trained on chest radiographs from multiple institutions, could partly predict which hospital an image came from—an astonishing feat that is entirely predictable once one notices that an institution's radiology plates, patient positioning, acuity mix, and disease prevalence leave signatures in pixels [18]. A model trained on one institution's data has no incentive to distinguish those signatures from pathology: within the training distribution, they correlate perfectly with disease, and the loss function rewards exploiting them. Performance only collapses when the correlation breaks—at a new institution, where the signature no longer tracks the disease—and the collapse is therefore guaranteed by the training setup, not by negligence [18]. The Epic sepsis audit supplies the deployed-system counterpart at commercial scale: the model's vendor-reported AUC of 0.76–0.83 became 0.63 under independent evaluation, with two-thirds of sepsis cases missed and a warning burden falling heavily on patients who never developed sepsis—a pattern the audit attributed partly to the model's reliance on features correlated with when sepsis was suspected rather than whether it was present [19]. Between them, the two studies define the validation gap's signature: a model that is excellent where it was made and unreliable where it is used, with no internal signal distinguishing the two regimes [18,19].

3.2. The Reliability Gap: Hallucination in Generated Medical Content

The second gap is generative. When an LLM fabricates a medical citation, the error is not cosmetic: it is a false representation of the clinical evidence base, with the burden of detection falling on time-pressed clinicians. The measured rates are sobering. Auditing ChatGPT-generated medical content, Bhattacharyya et al. found that 47% of the 115 generated references did not exist and a further 46% existed but were inaccurate—only about 7% were authentic and accurate [21]. The Med-HALT benchmark extended the question internationally, testing memory and reasoning hallucination across examination questions from multiple countries and showing that models fabricate plausible-but-false clinical knowledge, not merely citations [22]. Hallucination improves with model generation but does not vanish: in bibliographic citation

tasks, fabrication fell from roughly 55% for GPT-3.5 to about 18% for GPT-4 [23]. Retrieval-augmented generation (RAG), which grounds generation in retrieved documents [24], narrows the fabrication channel but does not close it—the model can still mischaracterize what a retrieved source says, and grounding introduces its own failure surface in retrieval quality. The residual risk is insidious precisely because grounding works: a grounded system's errors arrive attached to real sources, so its fabrications wear the appearance of scholarship. Detection, moreover, is professionally asymmetric—checking whether a citation exists is mechanical, but checking whether a cited study supports the claim made of it requires exactly the expertise that automation was supposed to economize.

The structure of the Bhattacharyya audit shows the failure mode with unusual clarity, because the authors graded every generated reference rather than sampling. Of 115 references produced across medical prompts, the 47% that did not exist at all were the detectable failure—a database query eliminates them—and the audit's deeper finding is the surrounding 46%: references that were real publications but misdescribed, cited for conclusions they did not reach, wrong in author, journal, or year [21]. This middle band is the industrially important one, because it passes the first check a busy clinician would run—the article exists, the title sounds right—and fails only the second, which requires reading the article. Med-HALT's cross-country design extends the same structure from citations to knowledge: using examination questions from multiple countries' medical curricula, it distinguishes memory hallucination (fabricating facts within the model's claimed competence) from reasoning hallucination (arriving at unsupported conclusions from given information), and finds both present across systems, with performance degrading on questions from regions whose curricula are underrepresented in training data [22]. The composite picture is a failure surface with a gradient: fabrication is densest where verification is hardest—nonexistent sources, misdescribed sources, underrepresented curricula—and thinnest where a checklist already exists. That gradient, not any single number, is the reliability problem: the residual errors are concentrated precisely where the professional's time is scarcest [21,22].

Legal medicine has an instructive parallel here: courts have sanctioned attorneys for filing fabricated AI-generated citations, and the profession's response—verification duties assigned to the licensed professional—is precisely the liability structure medicine is now improvising. The difference is that a fabricated medical citation can mislead treatment decisions for patients who never consented to be downstream of a language model.

3.3. The Utility Gap: From Accuracy to Outcomes

The third gap is the one patients actually experience: technical accuracy is not clinical benefit. Tschandl et al. demonstrated the asymmetry with an interventional study in skin-cancer recognition: AI assistance improved less experienced clinicians but degraded the diagnostic accuracy of experts, who did worse with the tool than either they alone or the AI alone [25].

The study's design deserves its reputation, because it measured collaboration rather than claiming it. Clinicians of graded experience diagnosed dermoscopic cases first unaided, then with AI suggestions presented alongside the image; the crossover structure let the analysis separate the AI's accuracy from the clinician's response to the AI—a distinction that aggregate "human-plus-AI accuracy" numbers, the era's dominant reporting format, structurally destroys [25]. The findings that survived that separation are the utility gap in miniature: experienced clinicians—whose unaided accuracy exceeded the AI's—were pulled toward the AI's errors when these appeared, a deference gradient invisible in any accuracy table; less experienced clinicians improved, meaning the same tool simultaneously raised the floor and lowered the ceiling of clinical performance [25]. The policy consequence is stark: the same deployment is beneficial or harmful depending on the user's expertise, so "does the tool work?" is not a property of the tool but of the tool–user–task triple, and no pre-deployment evaluation that fixes one user population can certify another. The collaboration surface has since acquired its own literature and its own vocabulary—automation bias, algorithmic aversion—but the 2020 study's core result remains the field's cleanest demonstration that the unit of evaluation for medical AI is not the model [25]. Human–AI collaboration is therefore not automatically superior to its components; it has its own performance surface that must be measured. More broadly, the clinical-impact literature has long identified the same deficiency: the overwhelming majority of reported AI performance comes from retrospective studies with no patient-level outcome, and evidence of improved outcomes remains the exception rather than the rule [6]. An AUC is a property of a dataset; mortality, time-to-treatment, and equity of access are properties of a health system. No benchmark score entails the second set. The corrective is outcome-linked evaluation—interventional studies in which the unit of analysis is the patient pathway—but such studies remain scarce relative to the deployment pace documented in Section 2.3. The scarcity has a structural cause as well as a financial one: outcome trials impose time constants—months to years of follow-up—that are alien to software release cycles, so deployment and evidence arrive on different clocks, and the system ships first. Medicine solved this misalignment for drugs by making the evidence clock the licensing clock; whether AI

deployment can be made to wait for its evidence is, at bottom, the regulatory question of Section 4.

Taken together, the three gaps explain the field's founding contradiction without paradox: the 2017 benchmark milestone [2] and the 2017 deployment failure [3] were measurements of different things, taken at different points of the translational pipeline, read as if they were the same.

4. Closing the Gaps: Governance as Clinical-Grade Evidence

The governance responses to the three gaps share a single logic: converting statistical performance into clinical-grade claims requires the same apparatus medicine built for therapies—tiered evaluation, independent audit, and post-market surveillance.

4.1. Regulation and Post-Market Surveillance

Regulatory practice is the most concrete instance. The U.S. FDA's list of AI-enabled medical devices reached 950 authorized products by August 2024, of which roughly 76% are radiological and the overwhelming majority were cleared through the 510(k) substantial-equivalence pathway rather than de-novo trials—a proportion that itself illustrates the validation gap: most authorized devices were never required to demonstrate performance on new, prospective data [26]. The EU Artificial Intelligence Act addresses the lifecycle rather than the entry event: AI systems that are safety components of medical devices fall into the high-risk tier, triggering risk management, clinical evaluation, human oversight, and—critically—post-market monitoring obligations, making deployed-model drift an auditable compliance object rather than a surprise [27].

The FDA list's composition tells the validation-gap story in administrative data. That roughly three-quarters of the 950 authorized AI-enabled devices are radiological is not a statement about where AI works best but about where evaluation is easiest: imaging has standardized inputs, established reference standards, and a regulatory pathway (510(k)) that certifies similarity to a predicate device rather than clinical performance on new data [26]. The pathway's logic—substantial equivalence—is inherited from devices whose behavior does not depend on a training distribution, and the growing share of adaptive, learning-enabled systems inside it is what pushed the regulator toward predetermined change-control protocols: documents that specify, before deployment, what model updates are permissible and what re-evaluation they trigger [26]. The DeepMind-era concern that motivated such thinking—models whose behavior is a function of data no one has enumerated—are, in regulatory terms, devices that keep

changing after the inspection; the change-control plan is the attempt to make continued learning compatible with continued accountability by converting model drift from an event into a documented process [26,27]. The machinery is young and its enforcement untested, but the direction is unambiguous: the permission-to-deploy conversation is moving from "was it validated?" to "who watches it, on what schedule, and what happens when it drifts?"—which is the validation gap being answered with governance instruments because the technical instruments did not arrive [26,27]. The WHO's guidance on ethics and governance of AI for health supplies the normative frame: transparency, inclusiveness, and accountability as design requirements, not aspirations [28]. LLMs stress this machinery: general-purpose chatbots do not fit device categories built for single-purpose instruments, and their deployment is running ahead of any settled regulatory pathway. A second tension is life-cycle: models that keep learning after authorization break the assumption, inherited from device regulation, that the validated artifact is frozen—hence the regulatory interest in predefined change-control plans that make continued learning compatible with continued accountability [26,27].

4.2. Equity as a Governance Obligation

The governance gap also has a distributional dimension. Obermeyer et al. dissected a widely used population-health algorithm and showed it systematically underreferred Black patients—not through any discriminatory variable, but because it used healthcare cost as a proxy for healthcare need, and Black patients historically received less spending for the same illness [29]. The bias was invisible in the model's accuracy metrics and emerged only when the algorithm's decisions were audited against outcomes.

Two features of the episode generalize beyond its case, and both sharpen the governance gap's definition. First, the mechanism was proxy selection, not data preprocessing: the algorithm predicted healthcare costs as a stand-in for healthcare needs—a choice that is individually defensible (costs are well-measured, complete, and predictive) and jointly discriminatory (equal spending does not mean equal illness across populations with unequal access) [29]. No fairness test applied at the feature level would have flagged it, because the proxy is not a "sensitive attribute"; the disparity lived in the validity of the substitution, which is invisible until the model's outputs are compared against an outcome the proxy was supposed to capture. Second, the episode's remediation shows what governance-grade correction looks like: the authors did not merely document the bias but rebuilt the proxy—reformulating the prediction target on illness measures collected directly from patients—which reduced the predicted disparity dramatically at similar accuracy [29]. Detect, re-specify, re-validate: that

loop is ordinary engineering once the audit exists, which locates the binding constraint exactly where this survey has placed it—not in the difficulty of fixing biased systems but in the institutional fact that nothing in the benchmark–clearance pipeline would ever have required the audit that revealed this one [29]. The lesson generalizes: fairness failures are not exotic failure modes but predictable consequences of proxy choices, undetectable at benchmark level and detectable only through governance-style audit. Equity testing before deployment, and surveillance after it, are therefore not additive virtues but load-bearing parts of the corrective for the utility gap.

4.3. Grading AI Claims with the EBM Hierarchy

The framework's synthesis is that the three gaps map onto medicine's existing evidence hierarchy. A retrospective benchmark is mechanistic-plausibility evidence; external and prospective validation is the analogue of a clinical trial; post-market surveillance is the analogue of pharmacovigilance; and structured reporting and audit are the analogue of trial registration and independent review [20,26,28]. Under this reading, the field's current turbulence is not a crisis of AI but a familiar stage of evidence maturation—the same stage at which therapies were once adopted on mechanistic enthusiasm before randomized evidence tamed adoption curves. The practical implication is that no single intervention closes all gaps: validation standards without outcome studies certify the wrong thing; regulation without equity audit entrenches disparity at scale; and none of it substitutes for clinicians who understand what the tools measure and what they do not [20].

The mapping deserves one more turn, because it converts a metaphor into a checklist. For each tier of the drug-evidence hierarchy, the AI analogue already exists in at least prototype form: trial registration maps to pre-deployment publication of evaluation protocols and data provenance, so that the evaluated configuration is the deployed one; randomized comparison maps to the interventional outcome studies of Section 3.3, including the collaboration designs that measure the tool–user–task triple; pharmacovigilance maps to the post-market monitoring obligations now entering force, whose artificial-intelligence-specific form is drift detection against declared performance; and independent review maps to the external audits—the sepsis evaluation [19], the bias dissection [29]—that demonstrated the model–record gap in both directions [20,26,28]. What the analogy also supplies is the expectation setting: nobody asks whether a drug "works"; one asks what evidence grade supports which indication at what dose. The AI field's interminable "does AI work in medicine?" debate dissolves under the same grammar—the answer is a ledger of claims, each with its evidence grade, its population, and its

surveillance plan [20,26,28]. The surveys reviewed in Section 1.3 catalogued the field's artifacts; the EBM mapping catalogs its evidence, and the difference between the two is the difference between a parts list and a pharmacopoeia.

5. Conclusion

This survey organized clinical AI along three translational gaps and reached three findings. First, measured capability has matured across three generations—specialist image classifiers, examination-scale language models exceeding 86% on licensing-style questions, and workflow-level deployments—with each generation shifting the measurement object from model to system. Second, three independent gaps—validation, utility, and governance—separate those measurements from patient benefit: retrospective scores degrade under dataset shift [18,19], generated content carries fabrication rates that decline but persist [21,23], and human–AI collaboration can underperform either party alone [25]. Third, the governance responses—risk-tiered regulation, post-market surveillance, and equity audit [26-29]—are converging on a single strategy with deep precedent: grading AI claims with the same evidence hierarchy evidence-based medicine applies to therapies.

The framework travels. Any domain combining statistical systems, licensed professionals, and high stakes—law and finance being the nearest—faces the same sequence: benchmarks arrive first, deployment exposes translational gaps, and evidence hierarchies plus surveillance close them. For medicine specifically, the highest-value open problems follow the gaps: funded infrastructure for external, prospective validation; outcome-linked studies of human–AI collaboration, whose asymmetric effects remain underexplored; standardized reporting of hallucination and grounding failure in clinical content; and a settled regulatory pathway for general-purpose medical LLMs. The distance from benchmark to bedside is real—but it is now mapped, and mapping it is the first step to closing it.

The mapping also explains what the next decade of the field will be about, because the three gaps define three distinct research programs rather than one. The validation program's frontier is shift-aware evaluation: benchmarks that sample deployment conditions rather than archives, and drift monitors that declare when measured performance has lapsed [18,20]. The utility program's frontier is the collaboration and outcome literature: study designs in which the unit of analysis is the tool–user–task triple [25], and endpoints are patient pathways rather than agreement statistics—a literature that will have to grow at the pace of clinical follow-up, not release cycles. The governance program's frontier is administrative: adapting evidence hierarchies built for molecules to systems that learn [26,27], and building the equity-audit

machinery that catches proxy harms before enrollment [29]. The survey's closing observation is that none of these programs is speculative—each has its founding results, cited above, and its institutional carriers already in motion. What remains is the unglamorous work of scale and enforcement, which is precisely the work medicine has done before, for every therapy now taken for granted. The benchmark-to-bedside gap is not a scandal to be lamented but a pipeline to be completed—and its completion is the field's actual task.

References

[1] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

[2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[3] Ross, C., & Swetlitz, I. (2017, September). IBM's Watson supercomputer recommended "unsafe and incorrect" treatments for cancer patients. STAT News.

[4] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38.

[5] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940.

[6] Kelly, C. J., Karthikesalingam, A., Suleyman, M., & Corrado, G. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.

[7] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410.

[8] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24, 1342-1350.

[9] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

[10] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 590-597).

[11] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.

[12] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker,

A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172-180.
[13] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Aguera y Arcas, B., … Natarajan, V. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.
[14] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.
[15] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198.
[16] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Webster, J. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, 1, 39.
[17] Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M. A., Kipnis, P., Liu, V., & Lee, K. (2024). Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst Innovations in Care Delivery, 5(3), CAT.23.0404.
[18] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683.
[19] Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065-1070.
[20] Finlayson, S. G., Subbaswamy, A., Singh, K., Fornage, J. F., Chodroff, L. C., Syrowatka, A., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. New England Journal of Medicine, 385(3), 283-286.
[21] Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus, 15(5), e39238.
[22] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-HALT: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343.
[23] Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 13, 14045.
[24] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33.
[25] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8), 1229-1234.
[26] U.S. Food and Drug Administration. (2024). Artificial intelligence-enabled medical devices. U.S. FDA. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
[27] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[28] World Health Organization. (2021). Ethics and governance of artificial intelligence for health. WHO.
[29] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

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

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

Yuying Zhang,Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: August 23 2026

Revised: August 30, 2026

Accepted: August 30, 2026

Published online: August 30, 2026

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

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

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

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

1. Introduction

1.1. Background and Motivation

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

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

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

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

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

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

1.2. Survey Methodology

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

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

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

1.3. Positioning Against Existing Surveys

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

Table 1 contrasts their coverage with this survey's.

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

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

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

2. Measured Capability: What Legal LLMs Can Do

2.1. From LEGAL-BERT to Legal LLMs

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

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

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

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

2.2. Benchmarks and What They Measure

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

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

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

Table 2. Major evaluation efforts for legal LLMs.

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

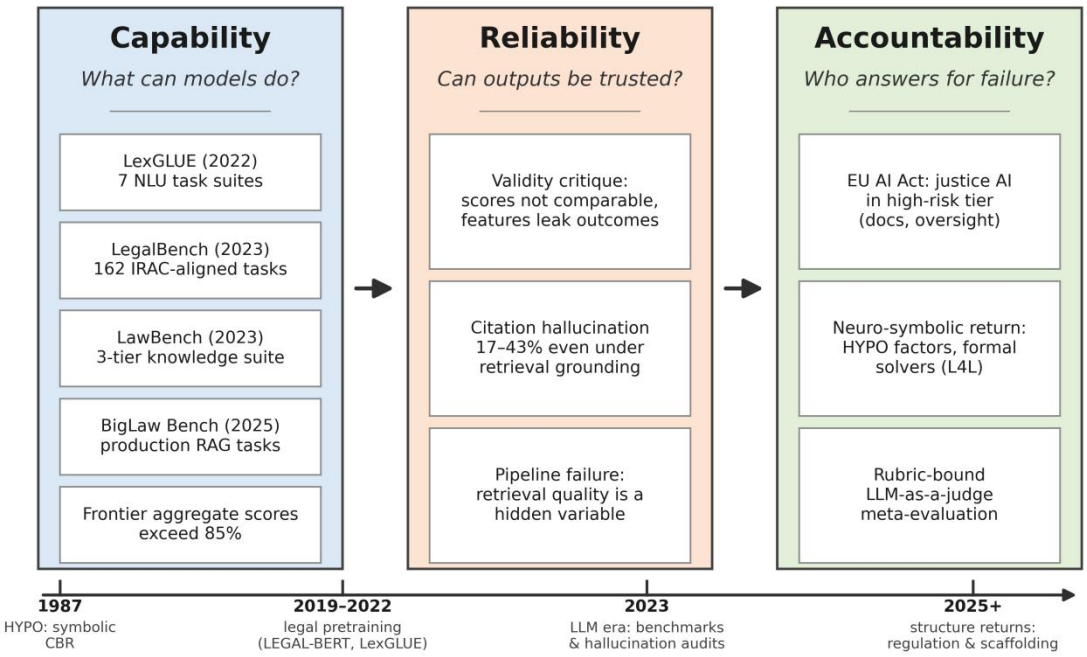

Figure 1. The capability–reliability–accountability (CRA) framework. Left: three generations of capability benchmarks. Middle: independent reliability failure sources. Right: accountability mechanisms that re-impose structure. Bottom: the field's timeline.

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

2.3. The Capability Boundary

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

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

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

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

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

3.1. The Evaluation Validity Critique

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

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

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

3.2. Hallucination and Fabricated Citations

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

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

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

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

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

3.3. Retrieval-Augmented Generation and Industry Benchmarks

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

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

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

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

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

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

4.1. Governance as External Structure

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

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

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

4.2. Neuro-Symbolic Scaffolding: Classical Legal Reasoning Returns

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

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

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

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

4.3. LLM-as-Judge Under Structural Scrutiny

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

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

5. Conclusion

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Impressum

Founders	Zhengjie Gao, Xinyu Song
Editor in Chief	Ao Feng, Chengdu University of Information Technology, China
Executive Editor	Zhengjie Gao, Geely University of China, China
Editorial Board	Jing Hu, Huazhong University of Science and Technology, China Xiaohu Du, Huazhong University of Science and Technology, China Xiangkui Li, Harbin University of Science and Technology, China Zuopeng Liu, Goettingen University, Germany Xinyu Song, Geely University of China, China
Young Editorial Board	Min Liao, Geely University of China, China Tao Zheng, Geely University of China, China Chong Li, Chongqing University, China Ruiqin Fan, Sehan University, Korea Ziyang Liu, Jiangsu Normal University, China Qiwei Liu, Urumqi Vocational University, China Minqiu Kuang, Hunan Agricultural University, China
Published By	**Hong Kong Dawn Clarity Press Limited** Rm 9042, 9/F, Block B Chung Mei Centre, 15-17 Hing Yip Street, Kwun Tong, Kowloon, Hong Kong e-mail: ijaaa@dawnclarity.press ***International Journal of Advanced AI Applications*** is published monthly.
Editorial Policy	***International Journal of Advanced AI Applications*** is directed to the international communities of scientific researchers in artificial intelligence, computers and electronic, from the universities, research units and industry.To differentiate from other similar journals, the editorial policy of IJAAA encourages the submission of original scientific papers that focus on the integration of the advanced AI applications. In particular, the following topics are expected to be addressed by authors: (1) Natural Language Processing (NLP): Conversational AI, machine translation, sentiment analysis, and context-aware dialogue systems. (2) Smart Cities and IoT Integration: AI for traffic optimization, energy management, waste reduction, and urban infrastructure. (3) Autonomous Systems and Robotics: Self-driving vehicles, drones, industrial automation, and human-robot collaboration.

	(4) Edge AI and Distributed Systems: Real-time processing, federated learning, and low-latency AI at the network edge. (5) Creative and Generative AI: Art, music, and content generation using generative adversarial networks (GANs) and transformers. (6) AI in Education and Industry: Adaptive learning platforms, intelligent tutoring systems, and AI-driven supply chain optimization. Ethical and Explainable AI (XAI): Fairness, transparency, and accountability in real-world AI deployment.