Stout in Law and Societal Studies (SLSS)

DOI: https://doi.org/10.61424/besja393
Journal Homepage: https://stoutjournals.org/index.php/SLSS

Retrieval Error or Generation Error? Hierarchical Failure Attribution and Evidence-Constrained Correction for Legal LLM-RAG

Adeline Zhou
Computer Science, University of Virginia, Charlottesville, VA, USA
Corresponding Author: Adeline Zhou, **E-mail**: azhou_uva@yahoo.com

ARTICLE INFORMATION

Submitted: April 05, 2026
Accepted: July 19, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Legal retrieval-augmented generation; legal information retrieval; failure attribution; evidence-constrained correction; citation support; selective prediction; Legal RAG Bench.

ABSTRACT

Legal retrieval-augmented generation (RAG) can fail because the supporting authority was not retrieved, because retrieved authority was not selected as evidence, or because selected evidence did not produce an adequate answer. Treating these outcomes as one accuracy score obscures the intervention required. This study evaluates those failure sources on Legal RAG Bench, comprising 4,876 passages from the *Victorian Criminal Charge Book* and 100 expert-written questions with long-form reference answers and passage-level relevance labels. Four retrievers—BM25, word TF-IDF, character TF-IDF, and reciprocal-rank fusion (RRF)—were crossed with three deterministic evidence-selection policies in 12 conditions, producing 1,200 question-condition observations. BM25 achieved the highest exact annotated-passage Hit@5 (0.34), whereas RRF paired with Query-Focused selection produced the highest answer token F1 (0.2308). Under a strict hierarchy applied to all observations, 74.25% were annotated-passage retrieval misses, 9.92% were evidence-selection failures, 2.58% were answer-realization failures, and 13.25% were strictly supported successes. An evidence-constrained correction gate increased mean token F1 from 0.2043 to 0.2184 and ROUGE-L from 0.1425 to 0.1580. The question-paired token-F1 improvement was statistically significant (Wilcoxon p = .0213; mean difference = 0.0141, 95% bootstrap CI [0.0031, 0.0267]), although the lexical citation-support proxy decreased slightly from 0.9493 to 0.9448. The lexical groundedness proxy equaled 1.0000 because every answer consisted only of sentences selected from its cited passages. In the lowest lexical-overlap quartile, every retriever had Hit@5 = 0, identifying vocabulary mismatch as the principal unresolved bottleneck. The results support component-specific legal RAG evaluation, retrieval-aware correction, and abstention policies that expose residual risk.

1. Introduction

Retrieval-augmented generation conditions an answer on external evidence rather than model parameters alone (Lewis et al., 2020). This architecture is attractive in legal practice because an answer can be connected to reviewable authority. Yet a citation-bearing answer is not necessarily correct, and a correct proposition is not necessarily supported by its cited authority. Studies have documented inaccurate legal claims in general-purpose language models (Dahl et al., 2024), while evaluations of commercial legal research systems show that retrieval augmentation reduces but does not eliminate hallucination (Magesh et al., 2025). The practical question is whether an error arose before evidence reached the answer interface, after appropriate evidence arrived, or where claims and citations were connected.

That question also has a governance dimension. Cross-border product counsel requires retrieval over multiple regulatory regimes, explicit source boundaries, and a record of which rule supports each recommendation (J. Li & Zhou, 2026), while multilingual humanitarian-information systems show that retrieval quality, calibration, and access equity can diverge even when an interface appears uniformly fluent (Y. Chen & Xu, 2026). A legal assistant must therefore treat evidence integrity and behavioral safety as coupled controls. Continual red-teaming, evidence-based self-verification, and multi-stage refusal research

each separates unsafe-input detection from response revision and utility preservation (Zheng et al., 2023; Zheng & Li, 2024; Zheng et al., 2024). Work on detecting and explaining dark patterns shows that seemingly innocuous interface language can itself become a risk-bearing part of the system (Xu et al., 2025). Human-facing risk cards further make prompt-injection and data-integrity concerns visible at the decision point rather than leaving them in an internal system log (Su, Rao, et al., 2026). These patterns motivate an evaluation that identifies the failing stage before prescribing a correction.

Existing evaluation traditions only partly answer this question. LegalBench measures a broad range of lawyer-designed legal reasoning tasks (Guha et al., 2023), and LexGLUE standardizes legal language-understanding tasks across heterogeneous sources (Chalkidis et al., 2022). JEC-QA demonstrates the combination of information retrieval and reasoning in examination-style legal questions (Zhong et al., 2020). Retrieval tracks such as AILA assess the ranking of cases or statutes for a factual scenario (Bhattacharya et al., 2019). These resources established important task-level baselines, but a pipeline that retrieves passages, synthesizes an answer, and attaches citations needs linked component measurements. A low end-to-end score can result from a missed authority, weak use of a retrieved authority, or unsupported wording. Each cause calls for a different remedy.

Legal RAG Bench was designed for this linkage. It contains 4,876 passages derived from the Judicial College of Victoria's Victorian Criminal Charge Book, together with 100 expert-crafted questions, long-form answers, and a specified supporting passage for each question (Butler & Butler, 2026; Judicial College of Victoria, n.d.). The questions present legally meaningful scenarios rather than isolated keyword lookups. The benchmark consequently permits the exact rank of an annotated passage to be connected to answer similarity and support measures. Its size also permits a complete factorial run over every question and passage without subsampling.

This study asks four questions. First, how accurately do sparse, word-level, character-level, and rank-fused retrievers recover the annotated passage? Second, how strongly do retrieval choice and evidence-selection policy affect answer similarity, lexical grounding, and citation support? Third, can a hierarchical rule distinguish retrieval, evidence-selection, and answer-realization failures, and can a confidence-triggered evidence correction improve the answer? Fourth, how does retained-answer risk change under selective coverage, particularly when the question and gold passage share little vocabulary?

The experiment makes three contributions. It first connects passage ranking and answer evaluation in a full 4 × 3 factorial design. It then applies a mutually exclusive hierarchy that tests exact annotated-passage retrieval, citation of that passage, and answer match in sequence. Finally, it evaluates a correction gate that expands retrieval from five passages to ten and applies RRF-based evidence fusion when any fixed confidence or citation trigger fires. Citation evaluation follows the broader movement from fluent-answer scoring toward explicit claim-source attribution (Gao et al., 2023; N. F. Liu et al., 2023).

The experimental scope is deliberately controlled. The three deterministic evidence-selection policies serve as transparent generator controls. Their answers are extractive combinations of retrieved sentences, which makes the mapping from ranking to evidence selection and citation fully inspectable. The reported lexical-grounding ceiling is therefore an architectural property, not evidence that open-ended legal generation is solved. Within that scope, the central result is clear: exact annotated-passage retrieval remains the dominant failure source, especially under low lexical overlap, while evidence-constrained correction yields a small but statistically reliable answer-similarity gain.

2. Literature Review

Legal NLP benchmarks differ in what they regard as success. LexGLUE covers tasks spanning legislation, cases, contracts, and human-rights decisions, showing that legal language is not one distribution (Chalkidis et al., 2022). LegalBench emphasizes lawyer-designed reasoning tasks (Guha et al., 2023), while JEC-QA combines legal knowledge and reasoning in multiple-choice examinations (Zhong et al., 2020). Expert-authored open questions instead require systems to find a relevant passage and express a responsive answer in terms that may differ from the source.

Recent resources focus more directly on retrieval. LegalBench-RAG uses carefully delimited relevant text spans to test whether retrievers can recover legal evidence without conflating retrieval with answer generation (Pipitone & Houir Alami, 2024). The Massive Legal Embedding Benchmark broadens embedding evaluation across legal tasks and jurisdictions, reinforcing the need for domain-specific testing rather than reliance on aggregate general-domain scores (Butler et al., 2025). Zheng et al. (2025) show that realistic legal research questions can have low lexical overlap with relevant material and may require reasoning-focused retrieval. That property is particularly important for hypothetical questions in which facts, doctrine, and procedural language are paraphrased. The present analysis retains retrieval-only measures but links them to downstream answer and citation outcomes.

The retrieval methods represent distinct assumptions. BM25 rewards query terms that are discriminative in the corpus while normalizing for document length (Robertson & Zaragoza, 2009). Word TF-IDF similarly relies on lexical alignment but represents unigram and bigram vectors under cosine similarity. Dense passage retrieval instead learns semantic representations that can reduce vocabulary mismatch (Karpukhin et al., 2020), while ColBERT retains token-level late interaction to balance semantic matching and fine-grained term evidence (Khattab & Zaharia, 2020). Cross-encoders can rerank an initial candidate set by jointly encoding a query and passage (Nogueira & Cho, 2019). The current experiment uses deterministic, locally reproducible lexical retrievers rather than trained dense or cross-encoder models, so the results establish a transparent floor and expose where learned semantic retrieval is most needed.

General benchmarks caution against assuming that one representation dominates across domains. BEIR documented zero-shot variability across retrieval collections (Thakur et al., 2021), and MTEB reached a similar conclusion across embedding tasks and languages (Muennighoff et al., 2023). Rank fusion combines incomparable component scores, but shared lexical blind spots can reinforce an error. Legal RAG Bench measures that possibility at the exact-passage level.

RAG evaluation has moved toward component-aware diagnosis. RAGAS separates context relevance, faithfulness, and answer relevance rather than reducing all behavior to one score (Es et al., 2024). ARES uses model-based judges and evaluates them against human annotations, highlighting the need to validate an evaluator instead of treating it as infallible (Saad-Falcon et al., 2024). RAGChecker decomposes retrieval and generation behavior into finer-grained indicators and shows how aggregate answer metrics can conceal opposing component effects (Ru et al., 2024). The hierarchy used here follows the same diagnostic principle but uses fully specified lexical and passage-identity rules, making every assignment traceable to stored question-level fields.

Citation quality is related to, but distinct from, answer quality. Gao et al. (2023) distinguish citation correctness and completeness in long-form answers, while N. F. Liu et al. (2023) show that generated search responses can appear informative without being fully verifiable. The Attributable to Identified Sources framework asks whether information is supported by an identifiable source (Rashkin et al., 2023). FActScore further motivates atomic evaluation because a long response can contain both supported and unsupported claims (Min et al., 2023). ExpertQA demonstrates the value of expert-curated questions and attributed long-form answers while also showing the cost of building reliable reference material (Malaviya et al., 2024). Together, this work supports the separate reporting of token overlap, sequence overlap, source containment, sentence-level support, and citation use.

2.1 Evidence Retrieval, Provenance, and Constrained Correction

Applied RAG studies outside law reinforce the value of separating retrieval from evidence use. Cloud-operations systems have paired private-document retrieval with hallucination-resistant question answering and bilingual incident summarization (B. Zhang et al., 2024; Liu et al., 2024), while evidence-chain designs explicitly connect word-level hallucination detection, source attribution, and provenance explanation (C. Li et al., 2024). Retrieval–summary integration for code intelligence similarly distinguishes whether a relevant artifact was found from whether its explanation was useful (Y. Li, 2024). Multi-hop retrieval work adds a budget dimension: an auditable controller must decide when another retrieval round is warranted, rather than assuming one top-k call is sufficient (Su et al., 2025). Query rewriting, hybrid retrieval, listwise reranking, and learned retrieval-quality prediction offer complementary ways to improve candidate ranking or identify weak rankings before generation (Meng et al., 2026; R. Zhang et al., 2025).

Several studies extend that decomposition into answer correction. Deterministic provenance explanations and unanswerable-question calibration treat abstention as an evidence-coverage decision rather than a stylistic refusal (Jin, 2025b; Zhong, Chen, et al., 2025). A command-safety checker in a DevOps copilot illustrates why retrieved text still requires action-level validation before release (B. Zhang et al., 2026). In finance and accounting, evidence retrieval, numerical guardrails, and evidence-constrained agents separately check source selection, arithmetic or temporal consistency, and the final decision trace (Y. Chen et al., 2025; Z. Li et al., 2026; S. Zhou et al., 2026). Scientific claim verification likewise ranks evidence before producing a narrative judgment (Su, Chen, et al., 2026). These systems address different domains, but they share a useful structural premise for legal RAG: correction should target the first failed dependency and should not silently trade source support for answer fluency.

The same principle appears in decision-support settings where a professional remains responsible for the outcome. A therapist-facing session copilot retrieves and ranks support from multi-turn dialogue rather than treating generation alone as the intervention (Y. Zhang & H. Zhang, 2025). Review-grounded recommendation evaluates whether an explanation is faithful to retrieved customer evidence, not only whether ranking quality improves (Chang et al., 2026). Trust-calibrated recommendation cards make this ranking–explanation distinction visible to users (B. Zhang et al., 2025). These analogues do not establish legal validity, but they clarify why a legal evaluation needs distinct measures for passage recovery, answer adequacy, citation support, and the decision to abstain.

2.2 Failure Attribution and Operational Diagnosis

Hierarchical diagnosis has close analogues in reliability engineering. Multi-source microservice analysis aligns metrics, logs, and traces before assigning a root cause (G. Liu et al., 2023), and noisy-neighbor modeling separates shared-infrastructure degradation from workload-specific residual behavior (Nie & Zheng, 2024). Retrieved normal prototypes provide a contrastive basis for explaining OpenStack failure-injection anomalies (Xin, 2025a), while retrieval-grounded HDFS analysis links detection, evidence bundles, narrative generation, and selective refusal (Sun et al., 2026). These pipelines demonstrate why an end-to-end error label is often insufficient: the observed failure can arise in sensing or retrieval, evidence fusion, classification or drafting, or the explanation layer.

More recent operational studies make the dependency chain explicit. Self-supervised log representation, conformal alert control, and cost-aware task routing address different stages of an incident workflow (C. Li et al., 2026; Xin, 2026f, 2026i). Generalist-agent trajectory prediction uses tool errors, replanning, loops, and recovery behavior to forecast success before manual evaluation (Zhong et al., 2026), while CloudTrail sequence modeling assigns events to interpretable attack-chain stages (Bai, Chen, et al., 2026). Intrusion studies similarly connect detection to language-guided features, suspicious-window localization, or operational mitigation rather than treating classification as the final product (Lu & Zhou, 2024; Y. Li & Lu, 2025; Wang et al., 2024; Xin, 2026d). For legal RAG, the analogous requirement is to preserve enough intermediate state to tell whether authority was absent, ignored, misapplied, or inadequately expressed.

2.3 Calibration, Selective Prediction, and Domain Shift

Calibration and selective action recur across high-cost decision systems. Calibrated uncertainty in heterogeneous capacity forecasting, uncertainty-driven rejection in credit scoring, and distribution-free uncertainty quantification in probability-of-default workflows all separate a point prediction from the decision to act on it (He et al., 2023; Jin et al., 2024; Y. Chen et al., 2023). Conformal demand envelopes extend this logic to resource oversubscription by controlling a risk boundary rather than optimizing only average error (S. Chen et al., 2024). The transferable lesson is not that a threshold from one domain applies to law, but that confidence must be evaluated against a declared loss and coverage policy.

Distribution shift makes that policy harder. Cross-cloud transfer and few-shot cold-start forecasting show why a model validated on established workloads may fail for a new topology or tenant (He et al., 2024, 2025). Workload-semantic forecasts and macro-market fusion similarly combine predictive accuracy with operational risk curves or regime-sensitive uncertainty (H. Zhou & K. Zhang, 2025; Zhao et al., 2025). Approximately shared features offer one formal route for transfer when domains overlap only partially, while uncertainty analysis for spectral estimators and rank aggregation clarifies that both the estimated representation and the combined ordering can be unstable (Zhong, Pan, et al., 2025; Zhong & Ling, 2024a, 2024b). In legal retrieval, jurisdictional vocabulary, procedural posture, and hypothetical phrasing can create the same broad problem: a confidence score learned in one region of the query distribution may not retain its meaning elsewhere.

Selective deployment studies also show several ways to surface residual risk. Calibrated resume screening, counterfactual credit explanations, and rubric-grounded essay scoring combine a predictive result with refusal, recourse, or human-review pathways (Jin, 2025a; Xin, 2026a, 2026c). Uncertainty-aware late fusion and object-level sensor fusion illustrate how confidence from several evidence channels can be calibrated before a fused decision is released (Xin, 2025c, 2026e). Conformal incident alerts and profit-aware admission control apply explicit gates when errors have asymmetric operational costs (Xin, 2026f; Zhao et al., 2026). Probabilistic demand forecasting and power-aware inventory planning further emphasize interval or tail-risk behavior rather than a single mean forecast (He et al., 2026; Xin, 2026g). The present study adopts the narrower implication that legal answers should be evaluated across coverage levels and that a correction trigger should be assessed for both benefit and harm.

2.4 Human-Facing Evidence and Oversight

An auditable pipeline also needs an interface that exposes its evidence state. Scientific-search cards organize retrieved sources, ranking confidence, and visual evidence hierarchy (Jin, 2025c); accounting review cards tie explanations to specific financial statements and notes (K. Zhang et al., 2025); and financial or infrastructure dashboards translate calibrated risk into a reviewable visual hierarchy (Liu et al., 2025; Z. Li et al., 2025). These designs treat provenance and uncertainty as information that a reviewer must be able to inspect, not merely as backend metrics.

Medical explanation and safety-card research makes the same distinction between prediction and communication. Visual explanations can be paired with calibrated uncertainty, while patient-facing cards can organize evidence, rubric results, and escalation guidance without implying that an interface itself validates clinical correctness (B. Zhou et al., 2025; C. Li et al., 2025; Zhong, Wu, et al., 2025). Incident visualization cards apply this idea to distributed logs by connecting evidence fragments to an actionable review surface (Nie et al., 2026). For legal RAG, the interface implication is modest but important: a reviewer should see the retrieved authority, the passages actually used, any support shortfall, and the reason for correction or abstention.

2.5 Privacy, Personalization, and Deployment Control

Evidence management becomes more difficult when a system retains user context. Confidence-gated memory writing and time-aware retrieval distinguish what should be stored, retrieved, updated, or forgotten across sessions (Sun et al., 2023), while federated topic-preference learning adds differential privacy to knowledge-grounded personalization (Kuo et al., 2025). Uplift modeling and cookieless incrementality estimation show why a recommendation policy should be evaluated against a counterfactual outcome and under reduced identifier availability (Bai, Wang, et al., 2026; Mu et al., 2023). Privacy-safe marketing-mix optimization offers an aggregate alternative when user-level identifiers are unavailable (Bai & Wu, 2026). In a legal setting, these studies support strict separation between matter-specific memory, retrieval relevance, and authorization to reuse prior context.

Personalization should also preserve an evidence path. Self-supervised customer representations can improve segmentation and next-action prediction, but a legal assistant cannot treat similarity as permission to transfer facts between users or matters (Xin, 2026h). Educational early-warning systems and strategy-aware emotional-support dialogue provide examples of routing uncertain cases to human intervention while retaining structured rationales or response-control labels (Xin, 2026b; Y. Zhang & Zhou, 2026). Finally, hybrid cloud–edge deployment highlights the interaction among routing confidence, privacy, cost, and model capability (Xin, 2025b). These concerns reinforce the need for the support-aware correction and held-out selective-risk analysis used below.

Legal applications make this separation consequential. When the annotated passage is retrieved but not cited, the outcome is an evidence-selection failure; when it is cited but token F1 remains below 0.25, the outcome is an answer-realization failure. A system can also retrieve only topically similar authority and accurately restate it; that answer may be lexically grounded yet unresponsive. Legal hallucination studies establish that persuasive form is a weak proxy for factual or doctrinal reliability (Dahl et al., 2024), and evidence from commercial tools confirms that citations must be inspected rather than counted (Magesh et al., 2025).

Finally, deployment requires an answer-or-abstain decision. Selective question answering ranks predictions by estimated reliability and reports risk as coverage changes (Kamath et al., 2020). Calibration research likewise shows that raw confidence is often misaligned with correctness (Guo et al., 2017). In legal RAG, selective evaluation should therefore use held-out confidence predictions and keep observations that share an annotated authority within one fold. This prevents the same gold passage from training and testing the confidence model through different questions or retriever–generator conditions. The present study applies this grouping and reports the observed risk–coverage curve without interpreting confidence as a guarantee of legal correctness.

3. Methodology

The experiment used the public Legal RAG Bench QA and corpus records (Isaacus, 2026). The QA file contained 100 newline-delimited JSON objects with an identifier, question, expert answer, and relevant_passage_id. The corpus contained 4,876 objects with an identifier, title, text, and footnotes. All 100 questions and all 4,876 passages entered every retrieval run. The files totaled 7,643,447 bytes. Their SHA-256 digests were e3b869a4e293d081ec5f5b39c2058c8d27b36f611aa9f8275eb1877a7c8b38b0 for QA and 3a3565bc5429f6cead90548e81f87352b449927e4be1cdd804c28d766bb9c246 for the corpus. Every annotated passage identifier occurred in the corpus.

Table 1 characterizes the data. The 100 questions referenced 95 distinct gold passage identifiers. Questions averaged 50.00 words, reference answers 50.31 words, and corpus passages 212.31 words; median passage length was 211 words. The corpus contained 218 normalized text repetitions beyond the first occurrence, including two annotated passages with duplicate-text counterparts under other identifiers. Exact-ID retrieval remained the primary measure. A duplicate-aware equivalent-hit check additionally counted any normalized text duplicate of the annotation as relevant; its Hit@5 values were identical to exact-ID Hit@5 for every retriever.

Table 1
Dataset characteristics and integrity checks

Measure	Value
Corpus passages	4876
Expert questions	100
Unique gold passage IDs	95
Question words, mean	50.0
Reference-answer words, mean	50.31
Passage words, mean	212.31
Passage words, median	211.0
Exact duplicate rows beyond first	218

Note. Counts and hashes were computed directly from the two benchmark files.

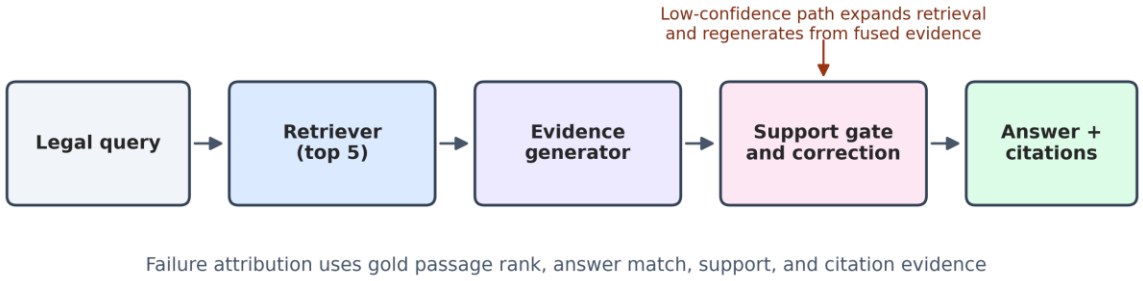

Figure 1. End-to-end evaluation, failure attribution, and evidence-constrained correction workflow.

Figure 1 shows the pipeline. Markdown syntax was removed, text was lowercased, and alphanumeric tokens were extracted. Content-token processing removed a fixed English stopword list and one-character terms. Titles and passage text were concatenated for indexing. Sentences were split at terminal punctuation, semicolons, or repeated whitespace; only sentences containing 5–120 content tokens were eligible. Figure 2 displays the empirical question, answer, and passage length distributions.

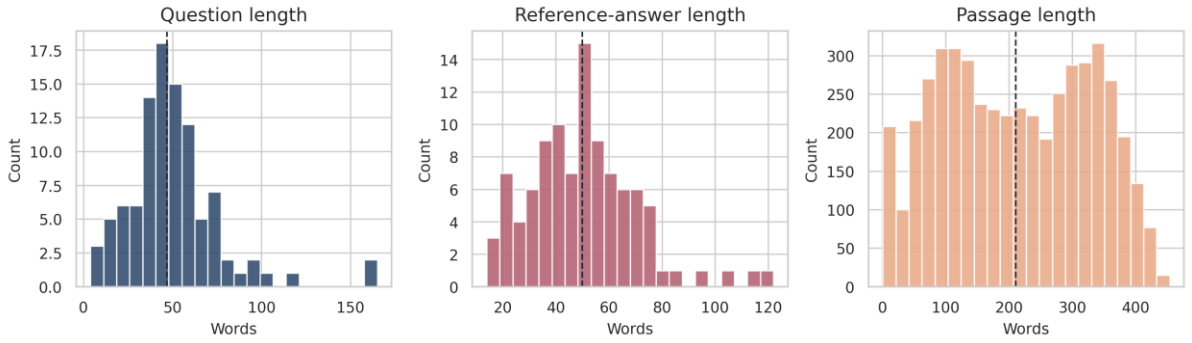

Figure 2. Observed word-count distributions for questions, expert answers, and corpus passages. Dashed lines mark medians.

Four retrieval systems ranked every corpus passage for every question. BM25 used k_1=1.5 and b=0.75 over content tokens. Word TF-IDF used word unigrams and bigrams, sublinear term frequency, L2 normalization, and a 120,000-feature cap. Character TF-IDF used within-word 3–5 character grams, a minimum document frequency of two, L2 normalization, and a 90,000-feature

cap. RRF combined the three complete rankings with reciprocal-rank constant 60. Stable sorting resolved tied scores. Generation used the top five passages; correction used the top ten.

Table 2
Fixed retrieval and evaluation configuration

Component	Setting
BM25	k1=1.5; b=0.75; lowercased content tokens; English stopword removal
Word TF-IDF	word 1-2 grams; sublinear TF; L2 norm; 120,000-feature cap
Character TF-IDF	character-within-word 3-5 grams; min_df=2; L2 norm; 90,000-feature cap
RRF Hybrid	BM25 + word TF-IDF + character TF-IDF; reciprocal-rank constant=60
Retrieval depth	k=5 for generation; k=10 for correction; retrieval reported at 1, 3, 5, and 10
Random seed	20260729
Bootstrap	2,000 question-level resamples; percentile 95% intervals

Note. Hyperparameters, retrieval depths, seed, and bootstrap procedure were held fixed across the full run.

Three evidence-selection policies formed the generator factor. Lead Evidence returned the first two eligible sentences in the top-ranked passage, capped at 80 words. Query-Focused scored each candidate sentence as $0.72O+0.20D+0.05/r+0.03L$, where O was IDF-weighted query overlap, D the min–max normalized document score, r passage rank, and L capped legal-cue density. It selected two sentences with pairwise token Jaccard similarity below 0.68 and capped output at 90 words. Evidence Fusion used $0.56O+0.22D+0.13C+0.06/r+0.03L$, where C was cross-document consensus among high-scoring sentences. It then subtracted 0.32 times maximum redundancy, retained up to three sentences with a positive selection margin above 0.02, and capped output at 100 words. Citations were the identifiers of passages supplying selected sentences.

Table 3
Generator controls, correction rule, and selective confidence

Component	Setting
Lead Evidence	first two eligible sentences from the top-ranked passage; 80-word cap
Query-Focused	two nonredundant sentences scored by query overlap, document score, rank, and legal cues; 90-word cap
Evidence Fusion	three MMR-filtered sentences using query overlap, document score, cross-document consensus, rank, and legal cues; 100-word cap
Correction gate	trigger when confidence<0.34, citation-support proxy<0.78, or citation-use proxy<0.75; replace with RRF top-10 evidence fusion
Selective confidence	five-fold group cross-validation by annotated gold passage; 300-tree random-forest regressor; depth=5; min_samples_leaf=6

Note. The gate applies RRF top-10 Evidence Fusion whenever any fixed trigger is met. Selective confidence is evaluated out of fold.

The 4 × 3 design created 12 conditions and 1,200 observations. The three deterministic evidence-selection policies serve as transparent generator controls: they hold answer construction inspectable while measuring how ranking and evidence use interact. Because generated words were copied from passages that were cited, the lexical groundedness proxy was 1 by construction. The citation-support proxy remained nontrivial because it measured how closely each answer sentence matched a sentence in the cited passages after truncation and sentence processing. The citation-use proxy measured the share of cited passages containing a sentence whose support score reached 0.25. Neither proxy is a semantic entailment judgment.

Table 4 defines the outcomes. Hit@k used exact passage identity. MRR@10 assigned reciprocal rank through ten and zero beyond ten; nDCG@10 used the single relevant passage's discounted position. Answer token F1 was the harmonic mean of unigram precision and recall after lowercasing and removing English articles. ROUGE-L was longest-common-subsequence F1. The lexical groundedness proxy was the share of answer content-token occurrences present in cited passages. The citation-support proxy was the mean best sentence-level token F1 between answer sentences and sentences in cited passages. Answer-match success required token F1 ≥ 0.25.

Table 4
Operational definitions of evaluation measures

Metric	Operational definition	Range
Hit@k	1 when the annotated supporting passage occurs in the first k results; otherwise 0	0-1
MRR@10	Reciprocal gold-passage rank through rank 10; zero outside the cutoff	0-1
nDCG@10	Single-relevance discounted gain of the annotated passage through rank 10	0-1
Token F1	Harmonic mean of unigram precision and recall against the expert answer after lowercase and article removal	0-1
ROUGE-L	Longest-common-subsequence F1 against the expert answer	0-1
Lexical groundedness proxy	Share of generated content-token occurrences present in the cited passages	0-1
Citation-support proxy	Mean best sentence-level token F1 between each answer sentence and sentences in its cited passages	0-1
Citation-use proxy	Share of cited passages with sentence-level support score of at least 0.25	0-1
Answer-match success	Token F1 of at least 0.25 against the expert answer	0/1
Selective lexical risk	One minus mean token F1 among answers retained at a confidence-ranked coverage level	0-1

Note. Support and grounding measures are deterministic lexical proxies rather than human legal judgments or natural-language-inference scores.

Hierarchical attribution assigned exactly one label per observation. If the exact annotated passage was absent from the top five, the label was retrieval miss. If the annotated passage was retrieved but its identifier did not appear among the selected citations, the label was evidence-selection failure. When the annotated passage was cited, token F1 below 0.25 produced answer-realization failure and token F1 at or above 0.25 produced strict supported success. The order separates retrieval availability from evidence use and then from answer adequacy. It is an operational diagnosis tied to the benchmark annotation, not a causal claim about every legally relevant passage; the corpus can contain additional useful passages that were not annotated. The hierarchy instantiates the component-separation principle used in fine-grained RAG diagnosis (Ru et al., 2024).

The correction gate used base confidence 0.52S+0.23M+0.25A, where S was the generator's selection confidence, M the top-score margin clipped to [0, 1], and A the mean top-five passage agreement with the other retrievers. It triggered when base confidence was below 0.34, the citation-support proxy was below 0.78, or the citation-use proxy was below 0.75. Whenever the gate triggered, the base answer was replaced by Evidence Fusion over the RRF top ten. Corrected confidence was 0.70F+0.20C+0.10U, where F was fusion confidence, C the citation-support proxy, and U the citation-use proxy. The gate did not use the expert answer or the gold passage identifier.

For selective analysis, a 300-tree random-forest regressor estimated held-out token F1, with maximum depth five, minimum leaf size six, and 75% feature sampling. Five-fold group cross-validation kept all observations sharing an annotated gold passage in the same fold, yielding 95 groups and preventing the five repeated-passage question pairs from crossing folds. Predictors included pipeline identity, confidence, the two citation proxies, answer length, and, for corrected outputs, gate and adoption indicators. Answers were ranked by these out-of-fold predictions, and selective lexical risk was computed at coverage from 0.40

to 1.00. This follows selective-QA practice (Kamath et al., 2020) while addressing confidence miscalibration concerns (Guo et al., 2017).

Percentile 95% confidence intervals used 2,000 bootstrap resamples. Condition-specific intervals resampled questions; aggregate base-versus-corrected intervals resampled the 100 question clusters while retaining all 12 conditions within each cluster. Friedman tests compared paired retrievers on Hit@5 and paired generator policies on question-averaged token F1. Two-sided Wilcoxon signed-rank tests evaluated generator pairs and question-averaged correction effects; generator-pair probabilities received Holm adjustment. The significance threshold was .05. All procedures used seed 20260729.

4. Results and Discussion

BM25 was the strongest exact-passage retriever. As Table 5 and Figure 3 show, its Hit@1, Hit@3, Hit@5, and Hit@10 were 0.15, 0.30, 0.34, and 0.39. Its MRR@10 was 0.2396 and nDCG@10 was 0.2764. RRF tied BM25 at Hit@1 (0.15) but reached only 0.26 at Hit@5; word and character TF-IDF reached 0.24 and 0.19. The retriever difference in paired Hit@5 outcomes was significant, chi-square(3)=16.4824, p = .0009. Duplicate-aware equivalent Hit@5 was identical to exact-ID Hit@5 for all retrievers, so the normalized duplicate groups did not change the primary top-five result. The broad bootstrap intervals reflect the 100-question sample, but the ordering and paired test agree that fusion did not improve annotated-passage recovery.

Table 5
Retrieval effectiveness on 100 expert questions

Retriever	Hit @1	Hit @3	Hit @5	Hit@5 95% CI	Equivalent Hit@5	Hit@10	MRR @10	nDCG @10	Median gold rank
BM25	0.150	0.300	0.340	[0.250, 0.430]	0.340	0.390	0.240	0.276	29.500
Word TF-IDF	0.120	0.200	0.240	[0.160, 0.320]	0.240	0.320	0.174	0.208	42.500
Character TF-IDF	0.120	0.160	0.190	[0.120, 0.270]	0.190	0.280	0.155	0.184	51.500
RRF Hybrid	0.150	0.240	0.260	[0.180, 0.350]	0.260	0.320	0.202	0.230	35.000

Note. Equivalent Hit@5 treats normalized duplicate passages as relevant. It was identical to exact-ID Hit@5 for all four retrievers.

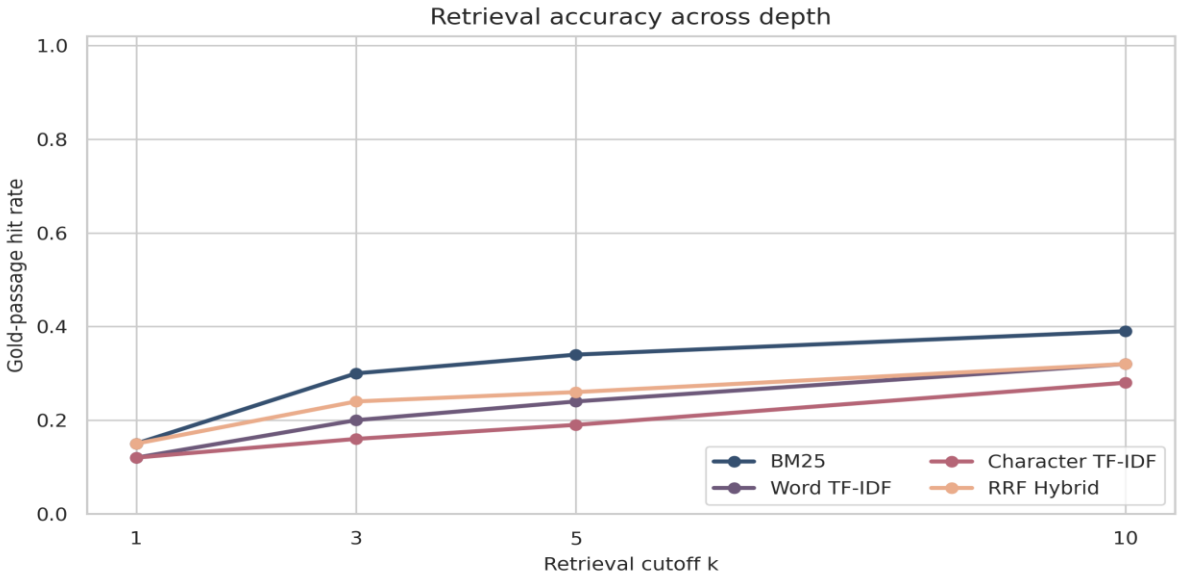

Figure 3. Gold-passage retrieval accuracy at cutoffs 1, 3, 5, and 10.

The answer results did not simply reproduce the retrieval ordering. RRF × Query-Focused achieved the best token F1, 0.2308 (95% CI [0.203, 0.261]), the best ROUGE-L, 0.1654, and the highest answer-match success rate, 0.37. BM25 × Query-Focused and BM25 × Evidence Fusion obtained token F1 values of 0.2090 and 0.2076 despite BM25's stronger Hit@5. Character TF-IDF × Query-Focused reached 0.2203 even though character retrieval had the lowest Hit@5. Figure 4 visualizes this interaction.

Table 6
Full retriever-by-generator factorial results

Retriever	Generator	Hit@5	Token F1	ROUGE-L	Citation-support proxy	Answer-match success
BM25	Lead Evidence	0.340	0.176	0.126	0.982	0.190
BM25	Query-Focused	0.340	0.209	0.136	0.940	0.300
BM25	Evidence Fusion	0.340	0.208	0.139	0.914	0.310
Word TF-IDF	Lead Evidence	0.240	0.171	0.124	0.975	0.160
Word TF-IDF	Query-Focused	0.240	0.206	0.144	0.927	0.290
Word TF-IDF	Evidence Fusion	0.240	0.202	0.143	0.926	0.340
Character TF-IDF	Lead Evidence	0.190	0.190	0.136	0.979	0.190
Character TF-IDF	Query-Focused	0.190	0.220	0.153	0.945	0.320
Character TF-IDF	Evidence Fusion	0.190	0.220	0.147	0.943	0.350
RRF Hybrid	Lead Evidence	0.260	0.196	0.139	0.974	0.210
RRF Hybrid	Query-Focused	0.260	0.231	0.165	0.954	0.370
RRF Hybrid	Evidence Fusion	0.260	0.223	0.157	0.933	0.360

Note. Answer-match success denotes token F1 of at least .25 against the expert answer. Generation uses the top five passages.

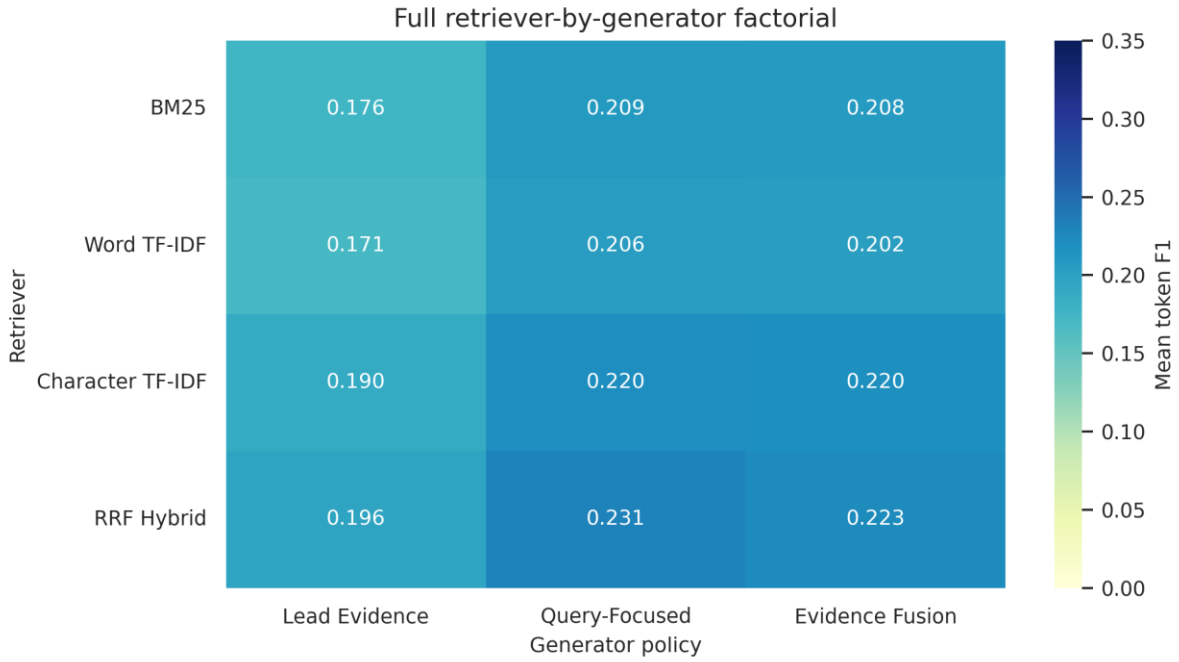

Figure 4. Mean token F1 for every retriever-by-generator combination.

This apparent reversal has two concrete explanations. First, answer metrics reward overlap with a long reference answer, not only recovery of one exact identifier. Top-ranked non-gold passages can contain relevant language, particularly in a corpus with 218

repeated texts. Second, RRF's top five can provide varied evidence even when the specifically annotated passage falls outside the cutoff. The benchmark's single passage label supports precise retrieval scoring, but it does not declare every other passage irrelevant. Consequently, Hit@5 and answer F1 are complementary rather than interchangeable.

Marginal results in Table 7 reinforce the generator effect. Query-Focused selection averaged 0.2166 token F1, Evidence Fusion 0.2132, and Lead Evidence 0.1832. The generator marginal range (0.0334) exceeded the retriever marginal range (0.0232). The generator Friedman test was significant, chi-square(2)=15.7834, p = .0004. Query-Focused and Evidence Fusion each exceeded Lead Evidence after Holm adjustment ($p < .001$), whereas they did not differ from one another ($p = .5383$). Query-Focused therefore gave the best balance: it captured responsive sentences without the additional answer length and cross-document material introduced by fusion.

Table 7
Marginal retriever and generator effects

Factor level	Token F1	ROUGE-L	Citation-support proxy
BM25	0.198	0.134	0.945
Word TF-IDF	0.193	0.137	0.943
Character TF-IDF	0.210	0.146	0.956
RRF Hybrid	0.216	0.154	0.953
Lead Evidence	0.183	0.131	0.977
Query-Focused	0.217	0.150	0.941
Evidence Fusion	0.213	0.147	0.929

Note. Values are means after averaging over the other experimental factor.

All conditions had a lexical groundedness proxy of 1.0000 because each answer consisted of content selected from the cited passages. This value does not establish substantive legal correctness. The citation-support proxy was highest for Lead Evidence (marginal mean 0.9773) because its answer used consecutive sentences from one cited passage. Query-Focused and Evidence Fusion traded a modest reduction in sentence-level lexical alignment for higher expert-answer overlap. That distinction accords with attribution frameworks in which source support and task responsiveness are separate dimensions (Rashkin et al., 2023).

Hierarchical attribution located most errors before answer construction. Across all 1,200 observations, 891 (74.25%) were exact annotated-passage retrieval misses, 119 (9.92%) were evidence-selection failures, 31 (2.58%) were answer-realization failures, and 159 (13.25%) were strict supported successes. Retrieval misses ranged from 66% to 81% across conditions, evidence-selection failures from 4% to 19%, answer-realization failures from 1% to 6%, and strict supported successes from 8% to 18%. RRF × Query-Focused and RRF × Evidence Fusion tied for the best composition: each produced 18 strict supported successes, 74 retrieval misses, seven evidence-selection failures, and one answer-realization failure. BM25 × Lead Evidence showed why gold-passage retrieval alone is insufficient: among its 34 questions with the annotation in the top five, 19 failed at evidence selection, six failed at answer realization, and nine achieved strict supported success.

Table 8
Hierarchical failure attribution by pipeline condition

Condition	Retrieval miss	Evidence-selection failure	Answer-realization failure	Strict supported success
BM25 × Evidence Fusion	0.660	0.140	0.040	0.160
BM25 × Lead Evidence	0.660	0.190	0.060	0.090
BM25 × Query-Focused	0.660	0.160	0.030	0.150
Character TF-IDF × Evidence Fusion	0.810	0.040	0.010	0.140

Condition	Retrieval miss	Evidence-selection failure	Answer-realization failure	Strict supported success
Character TF-IDF × Lead Evidence	0.810	0.070	0.040	0.080
Character TF-IDF × Query-Focused	0.810	0.040	0.010	0.140
RRF Hybrid × Evidence Fusion	0.740	0.070	0.010	0.180
RRF Hybrid × Lead Evidence	0.740	0.110	0.050	0.100
RRF Hybrid × Query-Focused	0.740	0.070	0.010	0.180
Word TF-IDF × Evidence Fusion	0.760	0.090	0.010	0.140
Word TF-IDF × Lead Evidence	0.760	0.120	0.030	0.090
Word TF-IDF × Query-Focused	0.760	0.090	0.010	0.140

Note. Each condition contains 100 questions. The hierarchy tests exact annotated-passage retrieval, annotated-passage citation, and reference-answer token match in that order.

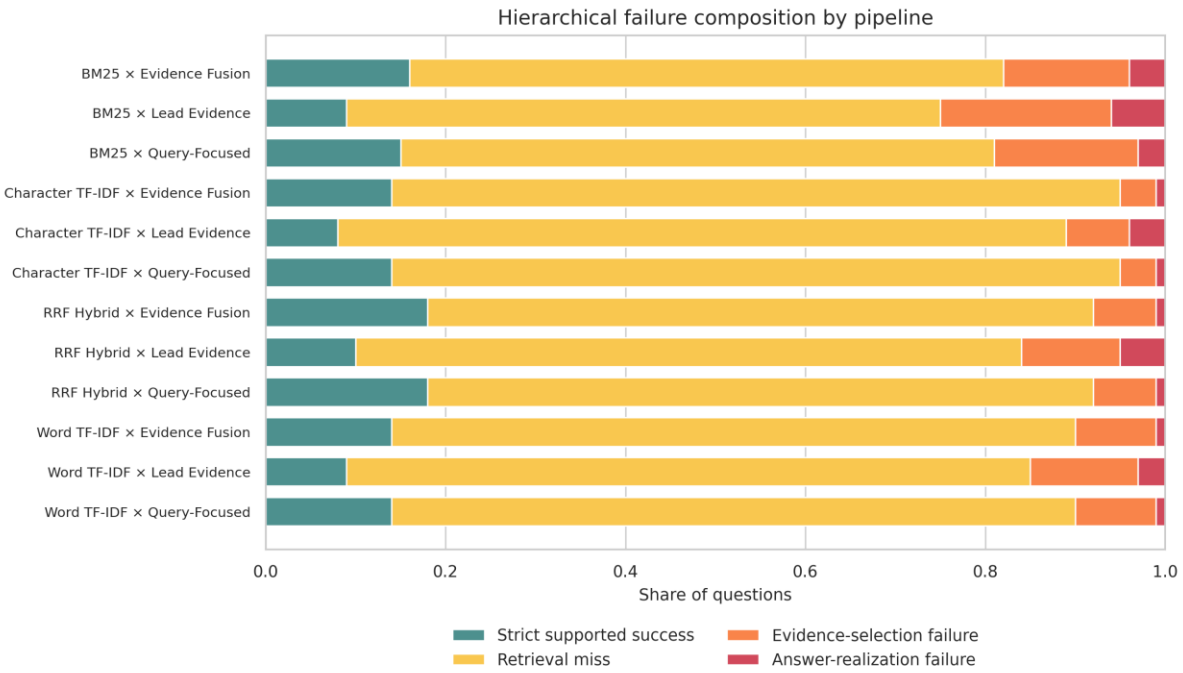

Figure 5. Hierarchical failure composition across the 12 factorial pipeline conditions.

The correction gate triggered and replaced the base answer in 64.25% of the 1,200 observations. Mean token F1 rose from 0.2043 to 0.2184, a gain of 0.0141 or 6.90% relative to the base mean. ROUGE-L rose from 0.1425 to 0.1580. When condition-level outcomes were averaged within each question, the paired token-F1 gain was significant ($p = .0213$), with a 95% question-cluster bootstrap interval of [0.0031, 0.0267]. The correction improved 416 observations, reduced token F1 in 316, and left 468 unchanged. Its mean effect was positive in eight factorial cells and negative in four; the largest negative cell was RRF × Query-Focused (−0.0032). The aggregate benefit therefore coexisted with condition-level harm.

Table 9
Aggregate effect of evidence-constrained correction

Stage	Token F1 [95% CI]	ROUGE-L [95% CI]	Lexical groundedness proxy	Citation-support proxy	Citation-use proxy	Repair adoption rate
Base	0.2043 [0.184, 0.225]	0.1425 [0.128, 0.158]	1.000	0.949	0.999	0.000
Corrected	0.2184 [0.192, 0.245]	0.1580 [0.137, 0.182]	1.000	0.945	0.996	0.642

Note. Bracketed intervals are 2,000-resample percentile intervals over question clusters; trigger and adoption rates coincide under the fixed replacement rule.

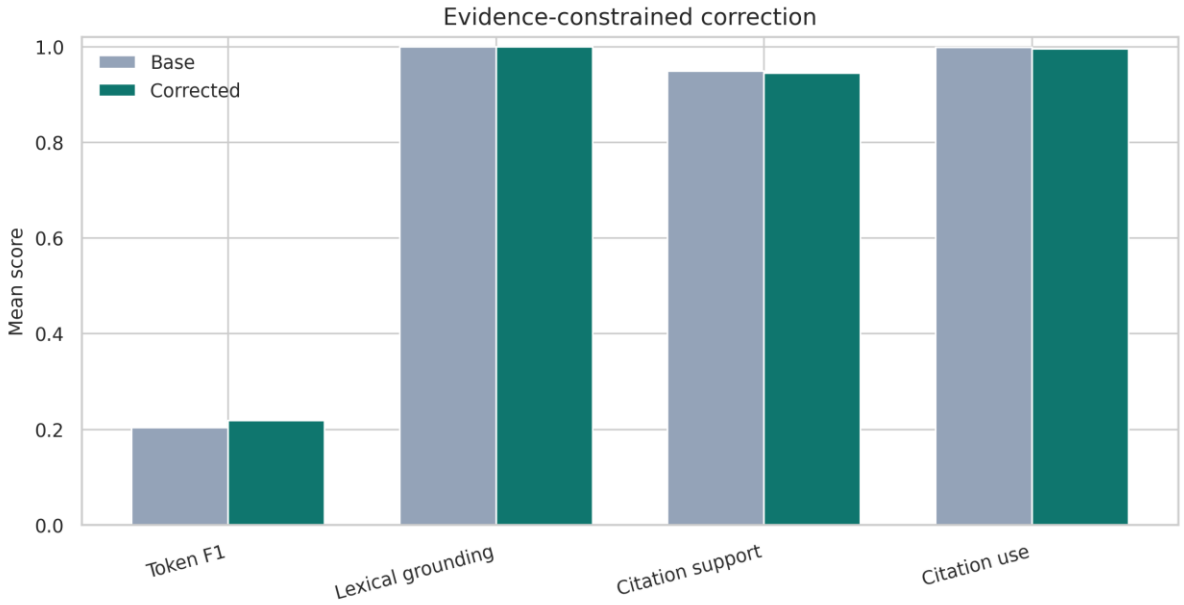

Figure 6. Mean answer and lexical citation-support measures before and after evidence-constrained correction.

The gain carried a small lexical-support trade-off. The citation-support proxy fell from 0.9493 to 0.9448, and the citation-use proxy fell from 0.9992 to 0.9963. Expanded fusion can add a responsive sentence from another passage, increasing overlap with the expert answer while weakening local sentence-to-citation alignment. The correction rule used fixed base-confidence and citation-proxy thresholds and had no access to reference answers or relevance labels at decision time. These results favor multi-objective gates with predeclared support floors and condition-specific validation. The finding also supports separate citation reporting advocated in prior generative-search evaluation (Gao et al., 2023; N. F. Liu et al., 2023).

Selective ranking reduced lexical risk most clearly at low coverage. At 40% coverage, corrected mean token F1 was 0.2557 compared with 0.2372 for the base output, producing selective lexical risks of 0.7443 and 0.7628. At 70% coverage, corrected and base risks were 0.7702 and 0.7799. At full coverage, correction reduced risk from 0.7957 to 0.7816. Both curves increased monotonically with coverage, indicating useful, though incomplete, out-of-fold ordering. Selective use improves the average retained answer, but the absolute lexical risks remain too high for autonomous legal advice.

Table 10
Confidence-ranked selective lexical risk

Stage	Coverage	Accepted responses	Mean Token F1	Selective lexical risk
Base	0.400	480	0.237	0.763
Base	0.500	600	0.232	0.768
Base	0.600	720	0.225	0.775
Base	0.700	840	0.220	0.780
Base	0.800	960	0.216	0.784
Base	0.900	1080	0.210	0.790
Base	1.000	1200	0.204	0.796
Corrected	0.400	480	0.256	0.744
Corrected	0.500	600	0.243	0.757
Corrected	0.600	720	0.235	0.765
Corrected	0.700	840	0.230	0.770
Corrected	0.800	960	0.225	0.775
Corrected	0.900	1080	0.221	0.779
Corrected	1.000	1200	0.218	0.782

Note. The confidence model used five-fold group cross-validation, keeping observations sharing an annotated gold passage in one fold.

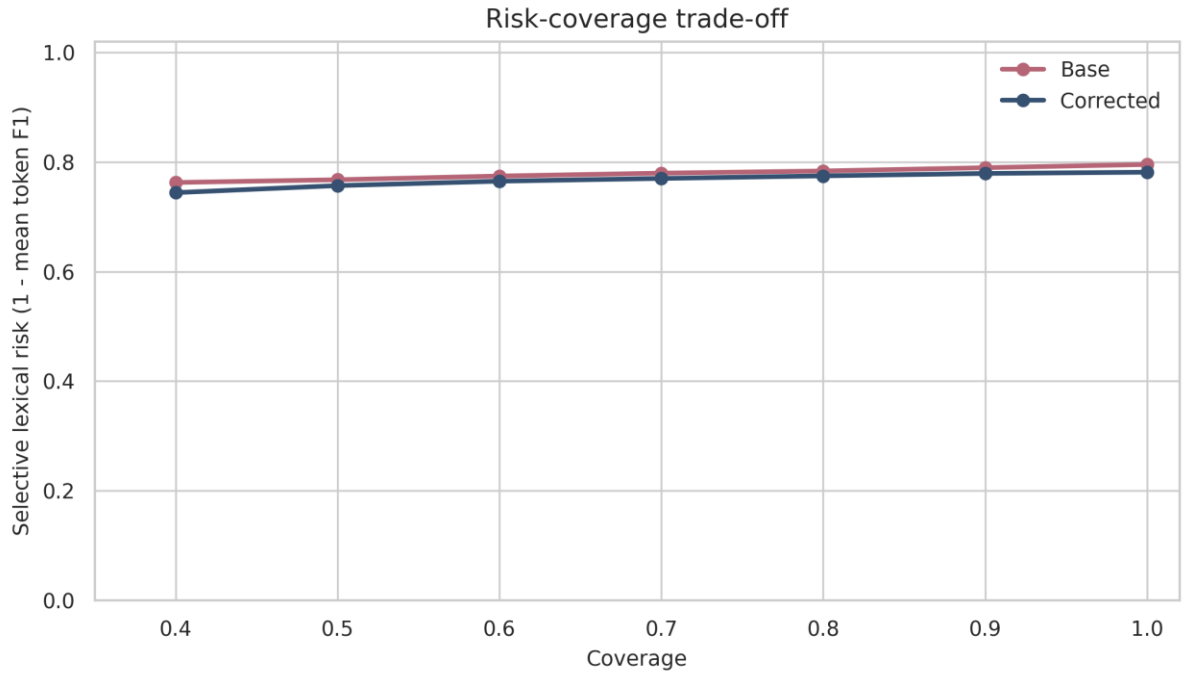

Figure 7. Selective lexical risk as coverage changes under gold-passage-grouped cross-validation.

Lexical overlap explained the most severe retrieval misses. Table 11 partitions questions into quartiles using IDF-weighted overlap between each question and its annotated passage. In the lowest-overlap quartile, mean overlap was 0.0502 and Hit@5

was 0 for BM25, word TF-IDF, character TF-IDF, and RRF. Median gold ranks ranged from 345 for RRF to 602 for character TF-IDF. In the highest quartile, mean overlap was 0.4444; BM25 Hit@5 reached 0.80 and the other retrievers reached 0.52–0.56. Figure 8 shows the inverse association between overlap and rank across individual questions.

Table 11
Retrieval sensitivity to question-evidence lexical overlap

Overlap quartile	Retriever	Questions	Mean overlap	Hit@5	Median gold rank
Q1 lowest	BM25	25	0.050	0.000	473.000
Q1 lowest	Word TF-IDF	25	0.050	0.000	350.000
Q1 lowest	Character TF-IDF	25	0.050	0.000	602.000
Q1 lowest	RRF Hybrid	25	0.050	0.000	345.000
Q2	BM25	25	0.119	0.160	48.000
Q2	Word TF-IDF	25	0.119	0.120	82.000
Q2	Character TF-IDF	25	0.119	0.040	72.000
Q2	RRF Hybrid	25	0.119	0.040	67.000
Q3	BM25	25	0.207	0.400	13.000
Q3	Word TF-IDF	25	0.207	0.280	19.000
Q3	Character TF-IDF	25	0.207	0.200	13.000
Q3	RRF Hybrid	25	0.207	0.440	8.000
Q4 highest	BM25	25	0.444	0.800	2.000
Q4 highest	Word TF-IDF	25	0.444	0.560	4.000
Q4 highest	Character TF-IDF	25	0.444	0.520	5.000
Q4 highest	RRF Hybrid	25	0.444	0.560	3.000

Note. Q1 contains the 25 questions with the lowest weighted lexical overlap; Q4 contains the highest-overlap questions.

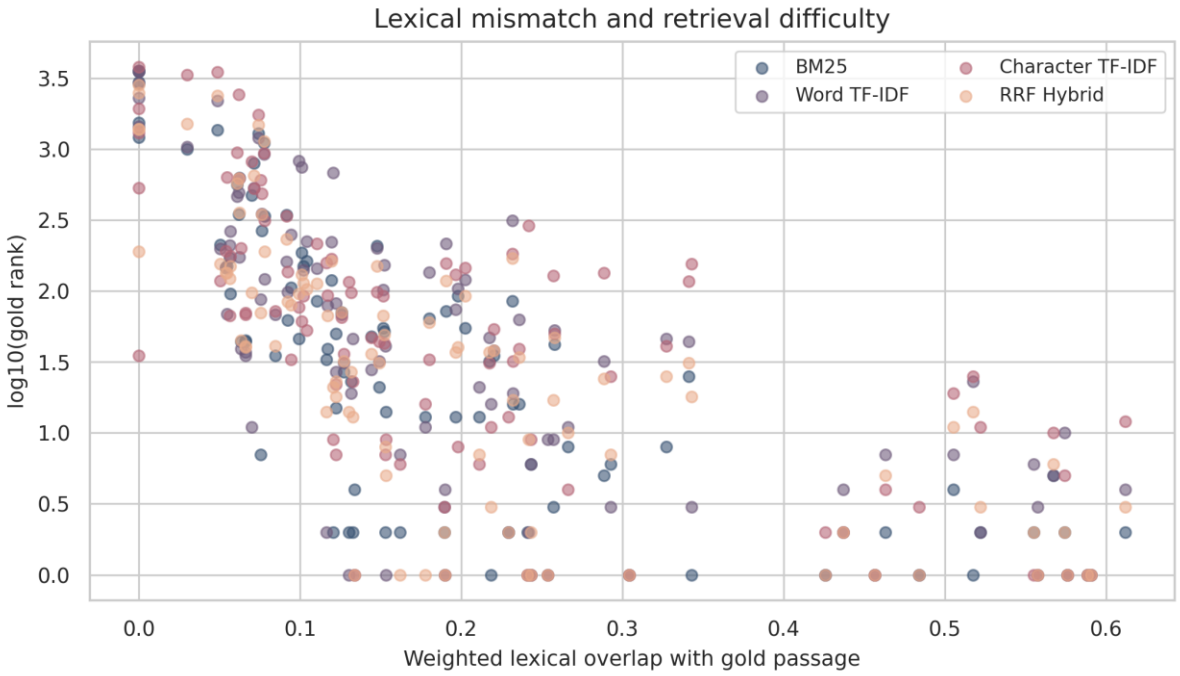

Figure 8. Relationship between weighted question-gold lexical overlap and gold-passage rank.

This quartile result is stronger than the aggregate comparison. Changing among four lexical or rank-fused retrievers did not recover a single gold passage in the hardest 25 questions at depth five. The next intervention should therefore change the information available to the query or ranking model, not merely adjust weights among the same signals. Reasoning-focused query expansion, learned legal embeddings, late interaction, or cross-encoder reranking are directly motivated by this failure mode (Karpukhin et al., 2020; Khattab & Zaharia, 2020; Zheng et al., 2025). Their benefit must still be measured on this legal collection because general embedding performance does not guarantee legal-domain ranking performance (Butler et al., 2025; Muennighoff et al., 2023).

Several limitations delimit the findings. The benchmark covers Victorian criminal law and procedure, so the measured rankings do not establish performance in contracts, tax, U.S. federal law, or multilingual practice. Single-ID scoring can penalize equivalent duplicate identifiers in principle, although equivalent Hit@5 and exact Hit@5 were identical in this run. Five passage identifiers each served two questions; grouping them in selective cross-validation reduced leakage, but passage-linked questions can still pose different factual distinctions. Token F1 and ROUGE-L measure lexical agreement with one expert answer and do not determine legal adequacy. The citation-support and citation-use proxies are local lexical-overlap measures, not judicial assessments or trained natural-language-inference judgments. Deterministic extractive selection isolates retrieval and evidence use, but it does not measure the reasoning, drafting quality, or hallucination rate of an open-ended LLM. Exact-ID misses are operational benchmark outcomes because useful unannotated passages may exist. These boundaries align the conclusions with the experiment.

5. Conclusions

The complete Legal RAG Bench evaluation separated retrieval, evidence selection, answer realization, citation support, correction, and selective lexical risk across 1,200 paired pipeline observations. BM25 retrieved the annotated passage most accurately at depth five, with Hit@5 = 0.34. End-to-end answer similarity followed a different ordering: RRF × Query-Focused achieved the best token F1 of 0.2308 and the highest looser answer-match success rate of 0.37. Under the strict hierarchy, retrieval misses accounted for 74.25% of observations, evidence-selection failures for 9.92%, answer-realization failures for 2.58%, and strict supported successes for 13.25%. RRF × Query-Focused and RRF × Evidence Fusion tied at 18% strict supported success within their respective conditions.

The evidence-constrained gate improved mean token F1 from 0.2043 to 0.2184, with a significant question-paired effect ($p = .0213$), and improved ROUGE-L from 0.1425 to 0.1580. The improvement was not free: the citation-support proxy decreased slightly from 0.9493 to 0.9448, and four of 12 factorial cells had negative mean token-F1 changes. The lexical groundedness proxy of 1.0000 resulted from the extractive, citation-bound construction and confirms source containment rather than general legal correctness. Selective ranking lowered average lexical risk at reduced coverage, but retained risk remained substantial.

The clearest bottleneck was lexical mismatch. All four retrievers had Hit@5 = 0 in the lowest-overlap quartile. Legal RAG development on this benchmark should prioritize semantic or reasoning-aware retrieval and evaluate it with the same passage-rank, answer-match, citation-support, and selective lexical-risk measures. More broadly, legal RAG systems should report where a failure occurred, apply a predeclared support-aware trigger, evaluate its answer–support trade-off on held-out data, and preserve human review whenever evidence recovery or answer adequacy remains uncertain.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Informatics, Electrical and Electronics Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923
[2] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment. *Journal of Technology Informatics and Engineering, 5*(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468
[3] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronics and Communications Systems, 6*(1), 83–95. https://doi.org/10.24042/ijecs.v6i1.30533
[4] Bhattacharya, P., Ghosh, K., Ghosh, S., Pal, A., Mehta, P., Bhattacharya, A., & Majumder, P. (2019). Overview of the FIRE 2019 AILA track: Artificial intelligence for legal assistance. In *Working notes of FIRE 2019—Forum for Information Retrieval Evaluation* (pp. 1–12). CEUR-WS.org. https://ceur-ws.org/Vol-2517/T1-1.pdf
[5] Butler, A.-R., & Butler, U. (2026). *Legal RAG Bench: An end-to-end benchmark for legal RAG* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.01710
[6] Butler, U., Butler, A.-R., & Malec, A. L. (2025). *The Massive Legal Embedding Benchmark (MLEB)* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2510.19365
[7] 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 (Volume 1: Long Papers)* (pp. 4310–4330). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.297
[8] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews. *JEECS (Journal of Electrical Engineering and Computer Sciences), 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2
[9] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/JACS.2024.40509
[10] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552
[11] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403
[12] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542
[13] 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. https://doi.org/10.1093/jla/laae003
[14] Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16
[15] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6465–6488). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398
[16] Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., . . . Li, Z. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems* (Vol. 36). https://proceedings.neurips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html
[17] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html
[18] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108
[19] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[20] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. *Journal of Technology Informatics and Engineering, 5*(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

[21] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[22] Isaacus. (2026). *Legal RAG Bench* [Data set]. Hugging Face. https://huggingface.co/datasets/isaacus/legal-rag-bench

[23] Jin, J. (2025a). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[24] Jin, J. (2025b). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[25] Jin, J. (2025c). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[26] Jin, J., Huang, T., & Lu, S. (2024). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[27] Judicial College of Victoria. (n.d.). *Victorian criminal charge book*. Retrieved July 29, 2026, from https://resources.judicialcollege.vic.edu.au/article/1053858

[28] Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5684–5696). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.503

[29] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

[30] Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 39–48). Association for Computing Machinery. https://doi.org/10.1145/3397271.3401075

[31] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[32] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[33] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[34] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[35] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[36] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[37] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[38] Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering, 4*(1), 284–305. https://doi.org/10.51903/jtie.v4i1.531

[39] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[40] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[41] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[42] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[43] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[44] Liu, N. F., Zhang, T., & Liang, P. (2023). Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7001–7025). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.467

[45] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72–87. https://doi.org/10.69987/JACS.2024.40809

[46] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies, 22*(2), 216–242. https://doi.org/10.1111/jels.12413

[47] Malaviya, C., Lee, S., Chen, S., Sieber, E., Yatskar, M., & Roth, D. (2024). ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3025–3045). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.167

[48] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[49] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741

[50] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience–creative–channel policies. *Journal of Advanced Computing Systems, 3*(1), 31–48. https://doi.org/10.69987/JACS.2023.30103

[51] Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2023). MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2014–2037). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.148

[52] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[53] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112–123. https://doi.org/10.69987/JACS.2024.40409

[54] Nogueira, R., & Cho, K. (2019). *Passage re-ranking with BERT* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1901.04085

[55] Pipitone, N., & Houir Alami, G. (2024). *LegalBench-RAG: A benchmark for retrieval-augmented generation in the legal domain* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.10343

[56] Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., & Reitter, D. (2023). Measuring attribution in natural language generation models. *Computational Linguistics, 49*(4), 777–840. https://doi.org/10.1162/coli_a_00486

[57] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

[58] Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Cheng, J., Wang, C., Sun, S., Li, H., Zhang, Z., Wang, B., Jiang, J., He, T., Wang, Z., Liu, P., Zhang, Y., & Zhang, Z. (2024). RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems* (Vol. 37, pp. 21999–22027). https://doi.org/10.52202/079017-0692

[59] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 338–354). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.20

[60] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[61] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[62] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[63] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[64] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computational Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[65] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (Vol. 1). https://openreview.net/forum?id=wCu6T5xFjeJ

[66] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. *Applied and Computational Engineering, 64*(1), 89–94. https://doi.org/10.54254/2755-2721/64/20241350

[67] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

[68] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195. https://doi.org/10.51519/journalisi.v7i3.1170

[69] Xin, Q. (2025c). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[70] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1–19. https://doi.org/10.66053/jaaie.v2i1.348

[71] Xin, Q. (2026b). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction. *Interdisciplinary Journal of Pedagogy and Research in Media Technology, 2*(1), 9–27. https://doi.org/10.64268/inspire.v2i1.117

[72] Xin, Q. (2026c). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German Credit Dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2), 215–231. https://doi.org/10.32664/j-intech.v14i02.2228

[73] Xin, Q. (2026d). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC), 8*(2), 325–334. https://doi.org/10.28989/avitec.v8i2.3973
[74] Xin, Q. (2026e). LiDAR–camera object-level fusion for multi-target tracking using JPDA and EKF: A reproducible empirical study on a PandaSet-parameterised five-sequence dataset. *Journal of Technology Informatics and Engineering, 5*(1), 54–76. https://doi.org/10.51903/jtie.v5i1.486
[75] Xin, Q. (2026f). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls (AVITEC), 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974
[76] Xin, Q. (2026g). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. *Findings*. https://doi.org/10.32866/001c.157499
[77] Xin, Q. (2026h). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. *Journal of Information and Technology, 14*(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229
[78] Xin, Q. (2026i). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *JEECS (Journal of Electrical Engineering and Computer Sciences), 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3
[79] Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. *Journal of Technology Informatics and Engineering, 4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491
[80] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308
[81] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697
[82] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534
[83] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710
[84] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). *Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.19481
[85] Zhang, Y., & Zhang, H. (2025). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547
[86] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. *Advances in Educational Technology and Psychology, 10*(2), 92–97. https://doi.org/10.23977/aetp.2026.100213
[87] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU Trace. *Journal of Technology Informatics and Engineering, 4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498
[88] Zhao, S., Ren, Y., & Chang, X. (2026). Profit-aware spot GPU admission control with cost-sensitive loss and evidence-grounded policy memos for AI workload supply-demand matching. *Journal of Technology Informatics and Engineering, 5*(2), 45–59. https://doi.org/10.51903/jtie.v5i2.545
[89] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/JACS.2024.40107
[90] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/JACS.2023.30203
[91] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106
[92] Zheng, L., Guha, N., Arifov, J., Zhang, S., Skreta, M., Manning, C. D., Henderson, P., & Ho, D. E. (2025). A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law* (pp. 169–193). Association for Computing Machinery. https://doi.org/10.1145/3709025.3712219
[93] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). JEC-QA: A legal-domain question answering dataset. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(5), 9701–9708. https://doi.org/10.1609/aaai.v34i05.6519
[94] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536
[95] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539
[96] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), Article iaae020. https://doi.org/10.1093/imaiai/iaae020
[97] Zhong, Z. S., & Ling, S. (2024b). *Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.05944
[98] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v258/zhong25a.html
[99] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[100] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[101] Zhou, H., & Zhang, K. (2025). News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015–2024 FRED panel. *Journal of Technology Informatics and Engineering, 4*(2), 487–501. https://doi.org/10.51903/jtie.v4i2.540

[102] Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. *Journal of Technology Informatics and Engineering, 5*(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544