Stout in Law and Societal Studies (SLSS)

DOI: https://doi.org/10.61424/ygk7h578
Journal Homepage: https://stoutjournals.org/index.php/SLSS

Reasoning-Focused Legal Retrieval-Augmented Question Answering under Low Lexical Overlap: Query Expansion, Evidence Citation, and Selective Abstention

Isabel Sun
Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
Corresponding Author: Isabel Sun, **E-mail**: isabel.sun_unc@gmail.com

ARTICLE INFORMATION

Submitted: March 22, 2026
Accepted: July 13, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Legal information retrieval; retrieval-augmented question answering; bar examination; BM25; latent semantic indexing; query expansion; citation evaluation; confidence calibration; selective abstention.

ABSTRACT

Legal retrieval-augmented question answering is difficult when a fact pattern and its controlling rule use different language. This study evaluates all 1,195 Bar Exam QA retrieval questions, 1,194 complete four-choice items, and the full 856,835-passage corpus. It compares BM25, 96-dimensional latent semantic indexing (LSI), reciprocal-rank fusion, pseudo-relevance feedback, answer-choice expansion, and cross-fitted rule-space expansion. LSI derives from TF–IDF and truncated singular-value decomposition; it is not a neural dense retriever. BM25 used a 40,000-term unigram/bigram vocabulary with k_1=1.2 and b=0.75. Mean query–gold TF–IDF cosine similarity was 0.0509, and 182 questions had zero overlap. Plain BM25 achieved 1.339% Recall@10; answer-choice expansion raised it to 5.105%, while rule-space BM25 reached 1.841%. The best retrieval-conditioned answer selector achieved 26.382% accuracy, versus 24.372% without evidence and 49.581% with gold evidence. Its 2.010-point gain over no evidence was nonsignificant (p=.265). Temperature scaling reduced its expected calibration error from 0.1486 to 0.0240, but risk–coverage ordering remained weak. Only 3.183% of depth-five contexts contained the gold or exact-equivalent passage, and 71.776% of questions combined missed gold evidence with a wrong answer. Stronger retrieval is therefore necessary before calibrated confidence can support dependable selective answering.

1. Introduction

Retrieval-augmented generation links language systems to inspectable sources (Lewis et al., 2020), which matters in law because narrow exceptions and authority can determine an answer. Yet general-purpose models can invent authorities (Dahl et al., 2024), and commercial legal research assistants still produce unsupported responses (Magesh et al., 2024). Legal QA must therefore retrieve and apply the rule, identify its source, and abstain when support is inadequate.

Bar-exam hypotheticals express concrete facts, whereas supporting passages state abstract rules and may omit salient fact terms. Bar Exam QA captures this gap with question–passage pairs collected through legal-research-like annotation (L. Zheng et al., 2025). BRIGHT similarly shows that reasoning-intensive retrieval exceeds surface matching (H. Su et al., 2025), while regulatory retrieval confirms that legally connected texts can be lexically dissimilar (Chalkidis et al., 2021).

The same separation between retrieval and justified action appears in other evidence-intensive systems. Cloud-operations RAG, bilingual incident triage, code intelligence, counseling support, financial search, multilingual humanitarian access, and cross-border product counsel all treat source acquisition as only one stage in a larger decision process (Y. Chen & Xu, 2026; J. Li & Zhou, 2026; Y. Li, 2024; Liu et al., 2024; Meng et al., 2026; B. Zhang et al., 2024; Y. Zhang & H. Zhang, 2025a). These studies do not establish performance on bar-exam law. They instead motivate the present separation of retrieval coverage, answer correctness, citation validity, calibration, and abstention.

Long-lived conversational memory and multi-source fault attribution further show that useful context must be selected, updated, and linked to a conclusion rather than merely accumulated (Liu et al., 2023; Sun et al., 2023). Evidence-chain methods make that linkage explicit through hallucination detection, provenance, retrieval-coverage analysis, and claim-level verification (Jin, 2025a;

C. Li et al., 2024; W. Su, Chen, et al., 2026; W. Su et al., 2025; Zhong, Chen, et al., 2025). Retrieval-grounded failure narratives and agent-trajectory analysis extend the same concern to operational decisions (Sun et al., 2026; Zhong et al., 2026), while retrieval-quality prediction asks whether ranking reliability can be estimated before generation (R. Zhang et al., 2025). The legal analogue is direct: authority must be located, identified, and shown to support the selected answer.

The experiment retains all 856,835 passages and 1,195 question–gold pairs; answer analysis uses 1,194 items because one lacks a fourth option. Transparent models compare BM25, low-rank LSI, reciprocal-rank fusion, feedback terms, cross-fitted projection toward supporting-rule vectors, and nonlearned answer-choice expansion.
The downstream component is an evidence-conditioned linear selector, not a large language model. Gold-evidence training and out-of-fold evaluation under no, gold, and retrieved evidence isolate retrieval from answer-selection limits.

Citation validity requires the gold identifier or an exact text hash at the cited depth, while support compatibility remains distinct from legal entailment. Evaluation includes fold-specific temperature scaling, NLL, Brier score, fixed and adaptive ECE, AURC, fixed coverage, and calibration-fold thresholds; probability alignment does not necessarily yield useful abstention (Guo et al., 2017; Si et al., 2022).

The study asks which retrieval formulation works best at low overlap, how evidence affects accuracy and attribution, and whether confidence isolates a reliable subset. Joint reporting prevents strength in one dimension from concealing failure in another.

2. Literature Review

2.1 Sparse, semantic, and fused retrieval

BM25 grew from probabilistic relevance weighting (Robertson & Spärck Jones, 1976) into a mature framework (Robertson & Zaragoza, 2009). It remains a strong zero-shot baseline across heterogeneous tasks (Thakur et al., 2021), including legal retrieval when queries and citations share discriminative terms (Rosa et al., 2021). Yet it cannot match an unnamed issue to a differently worded rule without useful overlap.

Neural dual encoders address mismatch: DPR improved open-domain QA retrieval (Karpukhin et al., 2020), E5 used weakly supervised contrastive pretraining (Wang et al., 2022), and later work improved instruction conditioning (Wang et al., 2024). In contrast, this study applies truncated singular-value decomposition to TF–IDF, testing corpus co-occurrence rather than a pretrained neural retriever.

RRF combines ranked lists without comparable scores (Cormack et al., 2009), but a weak component can add distractors when compressed vectors capture broad topics rather than narrow rules.

Cross-domain fusion research reinforces the need to validate each component rather than presume that combination is beneficial. Residual fusion for virtual-machine degradation, uncertainty-aware late fusion for perception, macro-market signal fusion, and LiDAR–camera tracking all make reliability contingent on the quality and calibration of the component signals (Nie & Zheng, 2024; Xin, 2025c, 2026d; H. Zhou & Zhang, 2025). Theoretical work on spectral ranking and its uncertainty likewise distinguishes an ordering rule from confidence in that ordering (Zhong & Ling, 2024a, 2024b). These are methodological analogues, not legal-retrieval results, but they support reporting fused and component rankings separately.

2.2 Query expansion for reasoning-intensive search

Query2doc improves sparse and semantic retrieval with hypothetical text (Wang et al., 2023); L. Gao et al. (2023) use HyDE to encode such text for zero-shot retrieval. Prompted expansion adds terms without retriever changes (Jagerman et al., 2023), corpus-steered expansion uses initially retrieved documents (Lei et al., 2024), and MILL verifies candidate expansions (Jia et al., 2024).

Expansion can also add the wrong concept. This study compares corpus-term feedback, cross-fitted ridge projection toward rule vectors, and nonlearned answer-choice expansion without a generative model.

Recent work also frames retrieval as a policy rather than a one-time term-generation step. Budgeted multi-hop retrieval chooses whether additional evidence is worth its cost, cost-aware AIOps routing chooses among parsing and diagnosis paths, and trajectory-reliability models anticipate tool-use failure (C. Li et al., 2026; W. Su et al., 2025; Zhong et al., 2026). Offline counterfactual evaluation and conservative policy selection provide related discipline for comparing decision rules without treating the most complex policy as automatically superior (Mu et al., 2025; Ye et al., 2026). Accordingly, the present expansions are compared at a fixed corpus scope and cutoff, with each added retrieval mechanism accountable for its own measured gain.

Reported BRIGHT and Bar Exam QA values are only contextual because preprocessing, queries, indices, and retrievers differ (H. Su et al., 2025; L. Zheng et al., 2025).

2.3 Legal reasoning, grounding, and citation quality

LegalBench shows that aggregate accuracy conceals distinct legal reasoning skills (Guha et al., 2023). Retrieval adds whether a system answers with authority, a split supported by Bar Exam QA (L. Zheng et al., 2025) and low-similarity regulatory retrieval (Chalkidis et al., 2021).

T. Gao et al. (2023) distinguish citation presence from correctness and completeness, and factual-consistency metrics require instance-level validation (Honovich et al., 2022). RAGAs separates faithfulness and relevance (Es et al., 2024); ARES builds task-specific judges (Saad-Falcon et al., 2024); FActScore checks atomic claims against sources (Min et al., 2023).

Applied systems increasingly make provenance visible to the reviewer. Accounting-aware retrieval treats documentary support as part of due diligence (Y. Chen et al., 2025), while evidence cards for scientific search and accounting disclosures expose source hierarchy and attribution (Jin, 2025b; K. Zhang et al., 2025). Evidence-constrained incident cards apply a similar pattern to distributed logs (Nie et al., 2026). These interfaces cannot establish that a legal conclusion follows from a passage, but they clarify why identifier-level citation, text equivalence, and answer support should be measured as distinct properties.

Here, citation correctness requires the gold identifier or identical normalized text. Lexical support remains diagnostic rather than legal judgment or entailment because legal systems can retrieve tangential material and produce unsupported answers (Magesh et al., 2024).

2.4 Calibration and selective answering

Temperature scaling can reduce ECE without changing predictions (Guo et al., 2017). QA probabilities still require held-out calibration (Jiang et al., 2021), and low ECE does not ensure that correct answers rank above errors (Si et al., 2022).

Calibration and selective intervention recur in other high-consequence settings. Infrastructure forecasting combines uncertainty estimates with capacity decisions, credit-risk workflows combine calibration with rejection or explanation, and conformal demand envelopes impose explicit operational bounds (S. Chen et al., 2024; Y. Chen et al., 2023; He et al., 2023; Jin et al., 2024a). Recruiter screening similarly couples calibrated matching with selective refusal, while operational risk curves expose the consequence of choosing a threshold (Zhao et al., 2025; B. Zhou, Jin, et al., 2025). These studies support calibration as a general design requirement, but their operating points do not transfer directly to legal QA.

Uncertainty communication is also inseparable from the downstream interface. Medical explanation and safety cards visualize uncertainty or risk for human review (C. Li et al., 2025; Ye et al., 2025; Zhong, Wu, et al., 2025), and counterfactual credit explanations add a contestable account of model decisions (Xin, 2026c). The relevant standard here is therefore not whether confidence looks well calibrated in aggregate, but whether lower-coverage decisions actually achieve lower realized legal-answer risk.

Selective QA answers or abstains, but maximum softmax probability may remain high on shifted errors (Kamath et al., 2020), and behavior varies across in-domain, shifted, and adversarial settings (Varshney et al., 2022). A legal policy therefore needs risk to fall as coverage decreases.

2.5 Distribution shift, deployment, and decision-policy evaluation

Distribution shift can invalidate a representation even when its source-domain fit appears adequate. Cross-cloud transfer and cold-start workload forecasting show how workload and topology changes alter prediction quality, while shared-feature theory formalizes when transfer across domains can remain possible (He et al., 2024, 2025; Zhong, Pan, et al., 2025). Few-shot medical pretraining, self-supervised customer representations, and review-grounded recommendation similarly examine how learned features travel to a new task or population (Chang et al., 2026; Mi et al., 2025; Xin, 2026e). For legal retrieval, the analogous shift lies between concrete facts and abstract rules and between benchmark passages and deployment-time authorities. The cross-fitted rule-space experiment tests one bounded version of that problem.

System constraints can also change which retrieval architecture is defensible. Retrieved normal prototypes and self-supervised log-anomaly models emphasize interpretable comparison to learned normality, while hybrid-cloud deployment makes latency and cost part of model selection (Xin, 2025a, 2025b, 2026f). Power-aware infrastructure planning joins forecasts to workload explanations rather than reporting accuracy in isolation (He et al., 2026). These analogues motivate the computational profile

reported here and caution against treating a retrieval improvement as operationally complete without accounting for evidence traceability and resource cost.

Rare failures require separate attention from average performance. Extreme-imbalance fraud and bankruptcy studies evaluate cost-sensitive learning and explanation under sparse adverse outcomes (Y. Chen et al., 2024; Jin et al., 2024b). Privacy-robust uplift estimation, conservative off-policy selection, and evidence-grounded admission control likewise distinguish predictive fit from the policy chosen from it (Bai, Wang, et al., 2026; Ye et al., 2026; Zhao et al., 2026). Those distinctions motivate paired tests, risk–coverage analysis, and explicit failure-cell counts rather than a single aggregate legal-QA score.

2.6 Adversarial robustness, privacy, and human oversight

Evidence access creates attack and privacy surfaces as well as retrieval opportunities. Continual red-teaming, evidence-based self-verification, and multi-stage refusal address adversarial prompts at the response-policy level (D. Zheng & Li, 2024; D. Zheng et al., 2023, 2024). Federated preference learning and identifier-loss studies address privacy at the data and measurement levels (Bai, Wang, et al., 2026; Bai & Wu, 2026; Kuo et al., 2025). Attack-chain detection, lightweight network intrusion detection, and language-guided feature selection illustrate complementary monitoring at the system layer (Bai, Chen, et al., 2026; Y. Li & Lu, 2025; Lu & Zhou, 2024). Privacy and data-integrity risk cards then make prompt-injection consequences reviewable by a human operator (W. Su, Rao, et al., 2026). Legal RAG requires the same layered view: provenance can be correct while the retrieved evidence, interface, or action remains unsafe.

Human oversight depends on presentation as well as model behavior. Dark-pattern detection warns that interface language can steer a user despite technically available information (Xu et al., 2025), whereas structural and visual consistency evaluation makes interface fidelity measurable (Y. Chen & Li, 2025). Scientific-search, medical-safety, e-commerce, accounting, financial-risk, infrastructure, counseling, and incident-response cards organize evidence and uncertainty for review (Jin, 2025b; C. Li et al., 2025; Z. Li et al., 2025; Liu et al., 2025; Nie et al., 2026; B. Zhang et al., 2025; K. Zhang et al., 2025; Y. Zhang & H. Zhang, 2025b; B. Zhou, Li, et al., 2025). Text-grounded design rationales and human-aligned design critique further emphasize that explanations should remain inspectable and editable (Y. Li et al., 2025; Wu et al., 2025). A legal evidence interface should therefore expose the passage, citation status, uncertainty, and abstention reason without implying that polished presentation validates the legal conclusion.

3. Methodology

3.1 Data, corpus integrity, and evaluation units

The question file contained 1,195 Bar Exam QA rows with fact patterns, questions, choices, answers, and gold passages. One row lacked choice D, so retrieval uses 1,195 questions and answer analysis 1,194 complete items. Answers were nearly balanced: 294 A, 295 B, 299 C, and 307 D. A disjoint-set procedure connected rows sharing a prompt or gold identifier, producing 889 leakage groups.

The corpus had 856,835 rows and unique identifiers: 847,228 case-law, 6,714 Wex, and 2,893 MBE passages. Text hashing found 838,017 unique texts and 18,818 duplicate rows. Length averaged 95.31 words (median 81; 95th percentile 224). Rows were retained intact, while retrieval features used at most 4,000 characters.

Table 1. Dataset scope and integrity checks

Component	Measured check	Count
Question records	Rows available for retrieval evaluation	1,195
Answer records	Items with four answer choices	1,194
Answer keys	Nonmissing answer labels	1,195
Gold evidence	Nonmissing passage texts and identifiers	1,195 / 1,195
Gold links	Unique gold identifiers resolved in corpus	1,149 / 1,149
Passage corpus	Rows and unique passage identifiers	856,835 / 856,835
Leakage control	Groups sharing prompt or gold passage	889
Missing choice D	Excluded from answer evaluation only	1
Missing subject label	Retained; not used as a feature	601
Answer balance	A / B / C / D	294 / 295 / 299 / 307

Note. Retrieval uses all 1,195 questions. Answer evaluation excludes only the item without choice D. Subject labels and corpus-source labels are not model features.

Figure 1 summarizes the experimental flow from indexed passages and grouped questions to retrieval, answer selection, calibration, citation checks, and error decomposition.

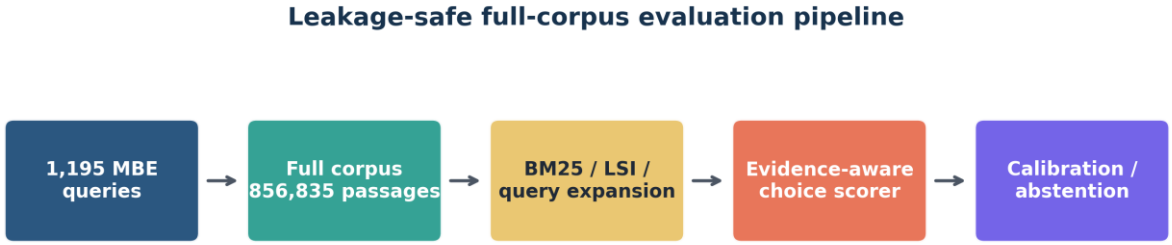

Figure 1. Experimental pipeline from the question and passage corpus through retrieval, evidence-conditioned answer selection, calibration, citation checks, and error decomposition.

3.2 Retrieval representations and systems

The BM25 vocabulary was inherited from a 40,000-feature TF–IDF model and consisted of English unigrams and bigrams after Unicode accent stripping, lowercasing, and English stop-word removal. Accordingly, BM25 and LSI vocabulary features excluded scikit-learn's English stop words. Only the separate tokenizer used for overlap analysis and expansion-term handling restored the designated legal function words, including "not," "must," "shall," and "unless." The BM25 index contained 36,195,706 nonzero weights, density 0.001056, and an average indexed document length of 53.532 terms. Parameters were k_1=1.2 and b=0.75, following the standard saturation and length-normalization formulation (Robertson & Zaragoza, 2009). Exact sparse matrix multiplication produced the top 100 corpus rows for each query.

LSI used the same maximum vocabulary size with word unigrams and bigrams, sublinear term frequency, L2 normalization, minimum document frequency 3, and maximum document frequency 0.995. Truncated randomized singular-value decomposition fitted 96 components on a deterministic stride-eight sample of 107,105 passages. The 96-dimensional vectors explained 0.08278 of sample TF–IDF variance. Every corpus passage was then transformed and L2-normalized. Similarity was the dot product of normalized vectors. Again, this is classical latent semantic indexing, not neural dense retrieval. Neural retrievers such as DPR and E5 are relevant prior work (Karpukhin et al., 2020; Wang et al., 2022), but no neural encoder supplied the LSI results.

Ten retrieval conditions were evaluated. BM25 and LSI used prompt plus question. BM25-AllChoices and LSI-AllChoices appended choices A–D. Hybrid-RRF and AllChoices-Hybrid fused corresponding BM25 and LSI lists using 1/(60+r) (Cormack et al., 2009). PRF-BM25 averaged TF–IDF weights from the top three initial BM25 passages, selected eight terms absent from the query, and appended them. RuleSpace expansion fitted ridge regression with α=10 from question LSI vectors to gold-passage LSI vectors. Five-fold StratifiedGroupKFold splitting, shuffled with seed 2025, kept leakage groups intact and approximately balanced answer labels. For each test fold, the mapping was fitted on the other folds, the projected vector generated RuleSpace-LSI results, and ten reconstructed terms with inverse-document frequency at least 1.5 were appended for RuleSpace-BM25. RuleSpace-Hybrid fused those two rankings.

Table 2. Fixed retrieval and expansion configurations

Component	Setting	Measured or fixed value
BM25	Index	40,000 word unigrams/bigrams
BM25	Scoring	k_1=1.2; b=0.75
BM25	Average indexed length	53.532 terms
LSI	Representation	TF–IDF + 96-dimensional SVD
LSI	Fit sample	stride 8; n=107,105
LSI	Explained variance	8.278%
AllChoices	Expansion	Prompt + question + choices A–D
RRF	Fusion	1 / (60 + rank)
PRF	Feedback	Top 3 BM25 passages; 8 new weighted terms
RuleSpace	Mapping	Ridge α=10; 10 reconstructed terms
RuleSpace	Validation	Five leakage-group-disjoint folds; seed 2025
All retrievers	Text cap and depth	4,000 characters; top 100

Note. The same corpus and top-100 evaluation depth are used for every retrieval condition.

3.3 Retrieval metrics and statistical analysis

For each question, strict relevance required the exact gold passage identifier. Equivalent relevance accepted any passage with the same normalized text hash. Recall was computed at 1, 5, 10, 20, 50, and 100; reciprocal rank was truncated at 100; nDCG used a single relevant item and cutoff 20. Query difficulty was measured by TF–IDF cosine similarity and token-set Jaccard similarity between the base query and gold passage. Questions were divided into four equal-sized overlap strata using ranked TF–IDF cosine values.

Three paired retrieval contrasts were prespecified: Hybrid-RRF versus BM25, RuleSpace-Hybrid versus Hybrid-RRF, and AllChoices-Hybrid versus Hybrid-RRF. Exact McNemar tests used discordant retrieval success at rank 10. Ninety-five percent confidence intervals for differences were obtained from 1,000 bootstrap samples of the 889 leakage groups, using seed 2025. This group bootstrap avoids allowing repeated prompts or shared gold passages to enter a resample as unrelated observations.

3.4 Evidence-conditioned answer selection

The answer experiment used the 1,194 complete items and the same retrieval folds. In each outer fold, one different fold served as a temperature-calibration set and the remaining three folds served as the fitting set. Word TF–IDF features used one- and two-grams with at most 30,000 features; character features used within-word three- to five-grams with at most 20,000 features. Vectorizers were fitted only on the fitting rows' questions, choices, and gold passages. A balanced logistic regression (C=1, liblinear solver, maximum 2,000 iterations, seed 2025) learned whether each question–choice pair represented the correct answer using gold evidence from fitting rows.

Nine features represented each candidate answer: word cosine between choice and evidence, character cosine between choice and evidence, query–choice word cosine, query–evidence cosine, cosine between evidence and the combined query-plus-proposed-outcome text, choice-token coverage by evidence, a negation-agreement indicator, log choice length, and log evidence length. The four candidate logits were normalized by softmax. Systems differed only in evidence: no passage; the gold passage; BM25 top 1; BM25-AllChoices top 1; or the first five BM25-AllChoices passages, each clipped to 900 characters and concatenated. Gold and single retrieved evidence were clipped to 4,000 characters.

Accuracy, macro-F1, NLL, multiclass Brier score, 15-bin ECE, 15-bin adaptive ECE, and standard AURC were calculated from out-of-fold predictions. Standard AURC ordered items by calibrated maximum softmax probability. A separate temperature was fitted for each system and fold by minimizing calibration-fold NLL over temperatures from 0.05 to 10. Because scaling does not alter the ordering of the four logits within an item, it changes confidence but not answer accuracy.

3.5 Citation and abstention evaluation

For retrieved-evidence systems, citation depth matched the supplied context: one for top-1 systems and five for the top-five system. Citation correctness used strict gold identifiers and equivalent text hashes. The logistic probability for the selected answer was also recorded as a support-compatibility proxy. Its discrimination was evaluated on 4,776 held-out gold-evidence

choice pairs, of which 1,194 were positive. ROC-AUC, average precision, and F1 at threshold 0.5 quantify the proxy itself; they do not convert it into an entailment label.

Selective risk equals one minus accuracy on answered items. For each predicted answer, a citation-aware selection score multiplied its calibrated maximum softmax probability by that answer's logistic support probability. Citation-aware AURC, the plotted risk–coverage curves, fixed-coverage analyses, and calibration-fold thresholds all used this product score; standard AURC alone used maximum softmax probability. Fixed-coverage analyses retained the top 100%, 80%, 60%, and 40% of product scores. Each calibration fold also supplied the largest prefix threshold whose observed risk was at most 10% or 20%, and that product-score threshold was applied to the associated test fold. The full design, metrics, and evidence conditions are reported in Table 3.

Table 3. Leakage-group-disjoint fold composition

Fold	Retrieval N	Answer test N	A	B	C	D	Reader fit N	Calibration N
0	237	237	58	58	60	61	718	239
1	240	239	59	59	60	62	715	240
2	240	240	59	59	60	62	714	240
3	240	240	59	60	60	61	716	238
4	238	238	59	59	59	61	719	237

Note. Retrieval-fold counts include the incomplete-choice item; reader test counts do not. Fit, calibration, and test partitions are disjoint within each outer fold.

4. Results and Discussion

4.1 Lexical-overlap profile

The measured query–gold TF–IDF cosine mean was 0.05095, with median 0.03554, first quartile 0.01294, and third quartile 0.07038. There were 182 zero-cosine pairs. Token Jaccard similarity was even smaller (mean 0.02853; median 0.024), with 159 zero-overlap pairs. Figure 2 shows concentration near zero. L. Zheng et al. (2025) reported a mean lexical-similarity estimate near 0.07 under their procedure; that prior value is not substituted for the 0.05095 observed here.

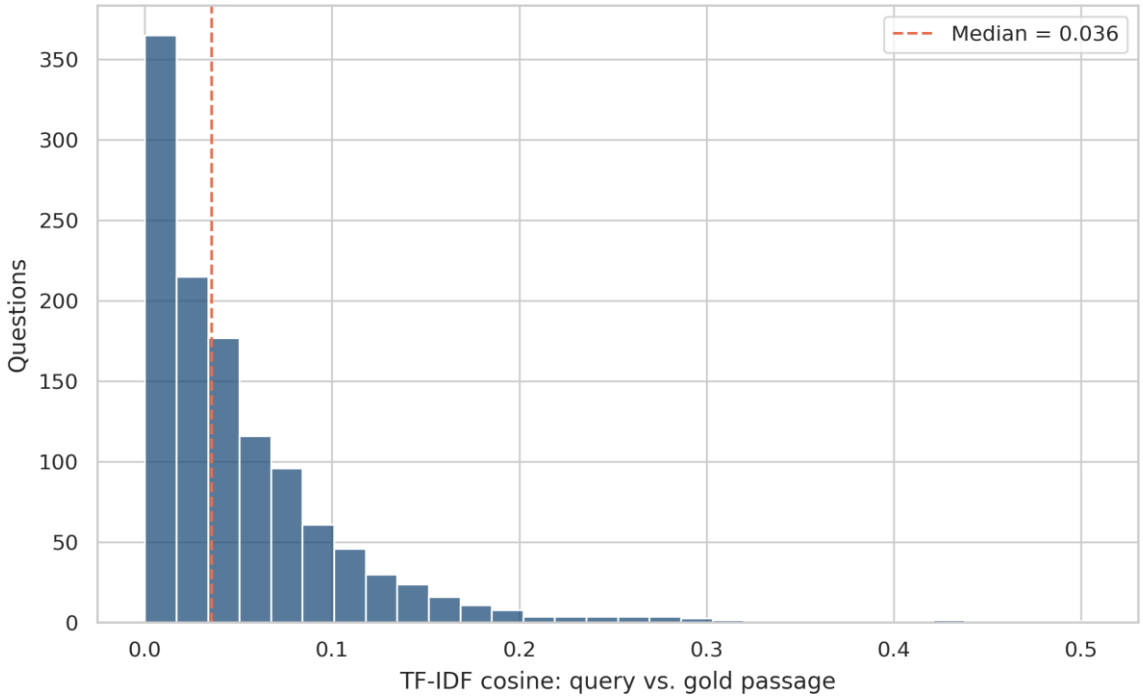

Figure 2. Distribution of TF–IDF cosine similarity between each question query and its annotated gold passage (N = 1,195).

Table 4. Corpus composition and measured lexical overlap

Panel A. Corpus composition

Domain	Measure	Value
Source	Case law	847,228
Source	Wex	6,714
Source	MBE	2,893
Text identity	Unique normalized hashes	838,017
Text identity	Duplicate-text rows	18,818
Passage length	Mean / median words	95.31 / 81
Passage length	95th percentile / maximum	224 / 44,644 words

Panel B. Question-to-gold overlap (N = 1,195)

Metric	Statistic	Value
TF–IDF cosine	Mean	0.05095
TF–IDF cosine	Median	0.03554
TF–IDF cosine	Q1 / Q3	0.01294 / 0.07038
TF–IDF cosine	Zero pairs	182
Token Jaccard	Mean	0.02853
Token Jaccard	Median	0.02400
Token Jaccard	Zero pairs	159

Note. TF–IDF cosine and token Jaccard values compare each base question query with its annotated gold passage under the experiment preprocessing.

4.2 Full-corpus retrieval performance

Table 5 reports all rank-based retrieval results. Plain BM25 obtained strict and equivalent Recall@10 of 0.01339, Recall@100 of 0.05105, MRR@100 of 0.00579, and nDCG@20 of 0.00826. Adding answer choices produced the strongest run: Recall@10 rose to 0.05105, Recall@100 to 0.17573, MRR@100 to 0.02227, and nDCG@20 to 0.03264. Thus, the choices multiplied Recall@10 by 3.81 and added 3.766 percentage points. Candidate outcomes supplied legal vocabulary that bridged facts and rules, although this formulation applies to multiple-choice rather than ordinary queries.

Table 5. Full-corpus equivalent-text retrieval performance

Method	R@1	R@5	R@10	R@20	R@100	MRR@100	nDCG@20
BM25	0.251%	0.669%	1.339%	2.008%	5.105%	0.0058	0.0083
BM25-AllChoices	0.753%	3.180%	5.105%	7.782%	17.573%	0.0223	0.0326
LSI	0.084%	0.084%	0.084%	0.251%	0.837%	0.0010	0.0012
LSI-AllChoices	0.000%	0.000%	0.251%	0.335%	1.841%	0.0006	0.0010
Hybrid-RRF	0.167%	0.251%	0.586%	1.255%	3.849%	0.0034	0.0049
AllChoices-Hybrid	0.502%	1.925%	3.013%	5.105%	13.138%	0.0139	0.0204
PRF-BM25	0.084%	0.753%	1.172%	1.506%	5.439%	0.0047	0.0063
RuleSpace-LSI	0.167%	0.586%	0.669%	1.339%	4.268%	0.0040	0.0056
RuleSpace-BM25	0.167%	1.004%	1.841%	2.594%	6.778%	0.0069	0.0103
RuleSpace-Hybrid	0.669%	1.004%	1.590%	2.762%	6.778%	0.0104	0.0133

Note. N = 1,195. Equivalent-text relevance accepts the gold identifier or a passage with the same normalized text hash. Percentages are shown for recall; MRR and nDCG remain on the 0–1 scale.

The LSI baseline was weak. Base LSI Recall@10 was 0.00084 and Recall@100 was 0.00837; LSI-AllChoices reached 0.00251 and 0.01841. The low 8.28% explained variance of the 96-component representation and the topical diversity of 856,835 passages

likely compressed away narrow distinctions. Consequently, equal-weight fusion did not help: Hybrid-RRF Recall@10 was 0.00586, 0.753 percentage points below BM25. AllChoices-Hybrid reached 0.03013, better than base hybrid but below BM25-AllChoices.

This result illustrates why fusion must be validated at the target decision boundary. Work on residual, market, perception, and sensor fusion conditions success on the information carried by each component rather than on the act of combination itself (Nie & Zheng, 2024; Xin, 2025c, 2026d; H. Zhou & Zhang, 2025). Here the LSI list contributed too little rule-level signal to offset the distractors it introduced, so a more elaborate ranking pipeline was measurably worse than its sparse component.

Expansion effects were mixed. PRF-BM25 obtained Recall@10 of 0.01172, slightly below BM25, even though Recall@100 rose from 0.05105 to 0.05439. Feedback sometimes moved a relevant passage into the top 100 while displacing it from the first 10, consistent with query drift from initially retrieved case facts. RuleSpace-BM25 reached Recall@10 of 0.01841 and Recall@100 of 0.06778, gains of 0.502 and 1.674 percentage points over BM25. RuleSpace-LSI reached Recall@10 of 0.00669. RuleSpace-Hybrid obtained Recall@10 of 0.01590 and had the highest Recall@1 among the non-choice-specific methods at 0.00669, but its Recall@100 equaled RuleSpace-BM25 rather than exceeding it.

Figure 3 displays recall curves across depths. Strict and equivalent aggregates were almost identical because duplicate-text alternatives rarely changed cutoff success. Both definitions remain important: the corpus contains 18,818 duplicate rows, and identifier-only evaluation can penalize identical text in other data arrangements.

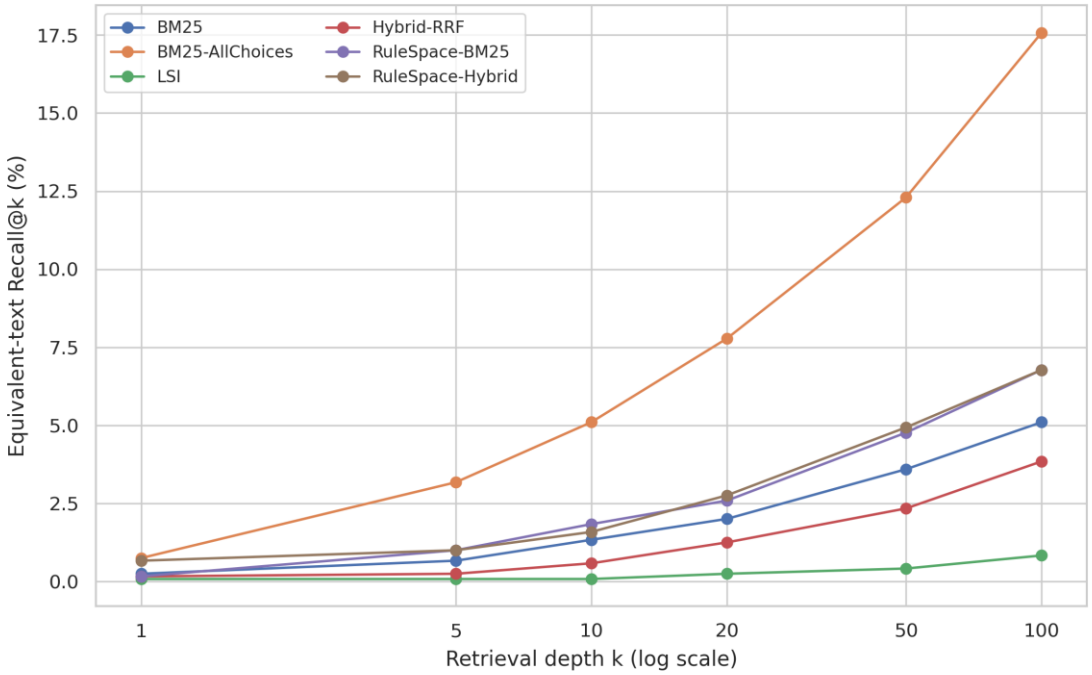

Figure 3. Equivalent-text Recall@k across representative full-corpus retrieval conditions.

L. Zheng et al. (2025) reported 5.03% BM25 Recall@10 under different retrieval and preprocessing; here, plain BM25 reached 1.339%, and only choice expansion reached 5.105%. Generative expansion results likewise are not estimates of the cross-fitted ridge run, whose RuleSpace-BM25 gain was 0.502 percentage points at rank 10.

4.3 Overlap sensitivity and paired tests

Retrieval varied sharply by overlap stratum. Plain BM25 retrieved no gold passage within 100 results in the lowest three quartiles at rank 10; its Q4 Recall@10 was 0.05351 and Recall@100 was 0.20401. BM25-AllChoices produced nonzero Recall@10 in every quartile: 0.01672, 0.01338, 0.05034, and 0.12375 from Q1 to Q4. This indicates that answer choices helped even when the base query and passage had almost no TF–IDF similarity. However, performance still rose more than sevenfold from Q1 to Q4.

Table 6. Overlap-stratified retrieval and paired statistical tests

Panel A. Equivalent recall by overlap quartile

Method	Q1 R@10	Q2 R@10	Q3 R@10	Q4 R@10	Q1 R@100	Q2 R@100	Q3 R@100	Q4 R@100
BM25	0.000%	0.000%	0.000%	5.351%	0.000%	0.000%	0.000%	20.401 %
BM25-AllChoices	1.672%	1.338%	5.034%	12.375 %	8.696%	9.030%	16.443 %	36.120 %
LSI	0.000%	0.000%	0.000%	0.334%	0.000%	0.000%	1.007%	2.341%
Hybrid-RRF	0.000%	0.000%	0.000%	2.341%	0.000%	0.000%	0.000%	15.385 %
RuleSpace-LSI	0.334%	1.003%	0.671%	0.669%	1.003%	6.355%	5.034%	4.682%
RuleSpace-BM25	0.000%	0.000%	0.671%	6.689%	0.334%	0.334%	2.685%	23.746 %
RuleSpace-Hybrid	0.334%	1.003%	0.336%	4.682%	1.003%	3.010%	4.698%	18.395 %

Panel B. Paired equivalent Recall@10 comparisons

Comparison	Difference	95% group-bootstrap CI	Discordant L:R	Exact p
Hybrid-RRF vs BM25	-0.0075	[-0.0127, -0.0025]	1:10	.0117
RuleSpace-Hybrid vs Hybrid-RRF	+0.0100	[+0.0034, +0.0175]	16:4	.0118
AllChoices-Hybrid vs Hybrid-RRF	+0.0243	[+0.0156, +0.0339]	31:2	1.31e−7

Note. Quartiles contain 298–299 questions. McNemar discordances are reported as left-only:right-only correct retrievals. Bootstrap resampling uses leakage groups.

Figure 4 contrasts selected methods across overlap quartiles. RuleSpace-LSI was the only base-query method with nonzero Recall@10 in Q1 (0.00334), and it peaked in Q2 at 0.01003 rather than Q4. RuleSpace-BM25 improved Q4 Recall@10 to 0.06689 but remained zero in Q1 and Q2. The cross-fitted mapping therefore supplied some semantic transfer while reconstructed terms mainly benefited questions already containing partial lexical cues.

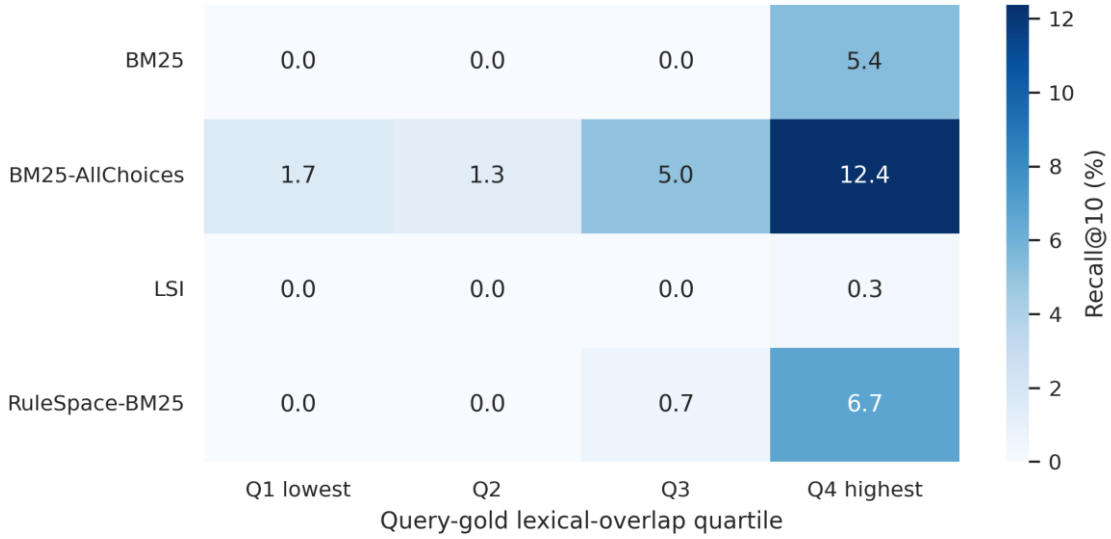

Figure 4. Equivalent Recall@10 by question-to-gold lexical-overlap quartile.

Paired tests confirmed that fusion and expansion effects were not interchangeable. Hybrid-RRF was 0.00753 below BM25 in equivalent Recall@10 (95% group-bootstrap CI [−0.01274, −0.00251]); discordances were 1 in favor of hybrid and 10 in favor of

BM25, with p=.0117. RuleSpace-Hybrid exceeded Hybrid-RRF by 0.01004 (95% CI [0.00338, 0.01753], p=.0118). AllChoices-Hybrid exceeded Hybrid-RRF by 0.02427 (95% CI [0.01563, 0.03387], $p=1.31\times10^{-7}$). These tests support a narrow conclusion: both supervised rule-space transformation and answer-choice text improved the weak base fusion, while the unmodified LSI component harmed BM25.

4.4 Answer selection and the evidence ceiling

Table 7 separates answer performance from retrieval. With no evidence, the selector achieved 0.24372 accuracy and 0.24357 macro-F1, close to the 25% four-choice chance level. Supplying the gold passage raised accuracy to 0.49581 and macro-F1 to 0.49571. This 25.21-percentage-point gap establishes that the feature-based selector could exploit a relevant passage, although it remained far from solving the legal reasoning task.

Table 7. Cross-fitted answer selection, calibration, and selective-risk metrics

System	Accuracy	Macro-F1	NLL	Brier	ECE	Adaptive ECE	AURC	Citation-aware AURC
No Evidence	24.372 %	24.357 %	1.3939	0.7538	0.0258	0.0744	0.8044	0.7908
Gold Passage	49.581 %	49.571 %	1.2064	0.6431	0.0723	0.0773	0.3926	0.3946
BM25 Top-1	26.298 %	26.285 %	1.3893	0.7519	0.0310	0.0433	0.7522	0.7413
BM25-AllChoices Top-1	24.539 %	24.534 %	1.3854	0.7496	0.0238	0.0412	0.7455	0.7401
BM25-AllChoices Top-5	26.382 %	26.381 %	1.3855	0.7496	0.0240	0.0397	0.7346	0.7278

Note. N = 1,194 complete-choice items. Accuracy and macro-F1 are percentages. NLL, Brier, ECE, and AURC use temperature-scaled probabilities.

Retrieved evidence delivered much smaller gains. BM25 top 1 reached 0.26298 accuracy. BM25-AllChoices top 1 reached 0.24539, despite its higher retrieval recall, while BM25-AllChoices top 5 reached the best retrieved-evidence accuracy of 0.26382. Against no evidence, the top-five gain was 0.02010. There were 225 questions correct only under top-five evidence and 201 correct only without evidence; exact McNemar p=.2651. Accordingly, this run does not establish a statistically reliable downstream accuracy improvement.

The mismatch reflects both low retrieval and sensitivity to distractors. Even the best retriever found the gold text in its top five for only 3.180% of retrieval questions, and the selector can change its answer when exposed to irrelevant evidence. AllChoices top 1 increased citation recall relative to base BM25 but produced nearly no accuracy gain over no evidence. Joint evaluation remains necessary (Es et al., 2024; Saad-Falcon et al., 2024).

4.5 Calibration and selective abstention

Temperature scaling materially reduced conventional calibration error for retrieved-evidence systems. For BM25 top 1, ECE fell from 0.08956 to 0.03099 and adaptive ECE from 0.10651 to 0.04326. For AllChoices top 1, ECE fell from 0.20028 to 0.02383; for top 5, it fell from 0.14855 to 0.02398. Top-five NLL declined from 1.48749 to 1.38549 and Brier score from 0.80282 to 0.74963. Figure 5 shows the raw and scaled reliability curves for the top-five condition.

Table 8. Calibration before and after temperature scaling

System	Raw NLL	Scaled NLL	Raw Brier	Scaled Brier	Raw ECE	Scaled ECE	Raw adaptive ECE	Scaled adaptive ECE
No Evidence	1.3929	1.3939	0.7534	0.7538	0.0400	0.0258	0.0513	0.0744
Gold Passage	1.2071	1.2064	0.6474	0.6431	0.0884	0.0723	0.0873	0.0773
BM25 Top-1	1.4334	1.3893	0.7750	0.7519	0.0896	0.0310	0.1065	0.0433
BM25-AllChoices Top-1	1.5465	1.3854	0.8310	0.7496	0.2003	0.0238	0.2003	0.0412
BM25-AllChoices Top-5	1.4875	1.3855	0.8028	0.7496	0.1486	0.0240	0.1532	0.0397

Note. Each outer test fold uses a temperature estimated on its associated calibration fold. Scaling does not change the predicted answer class.

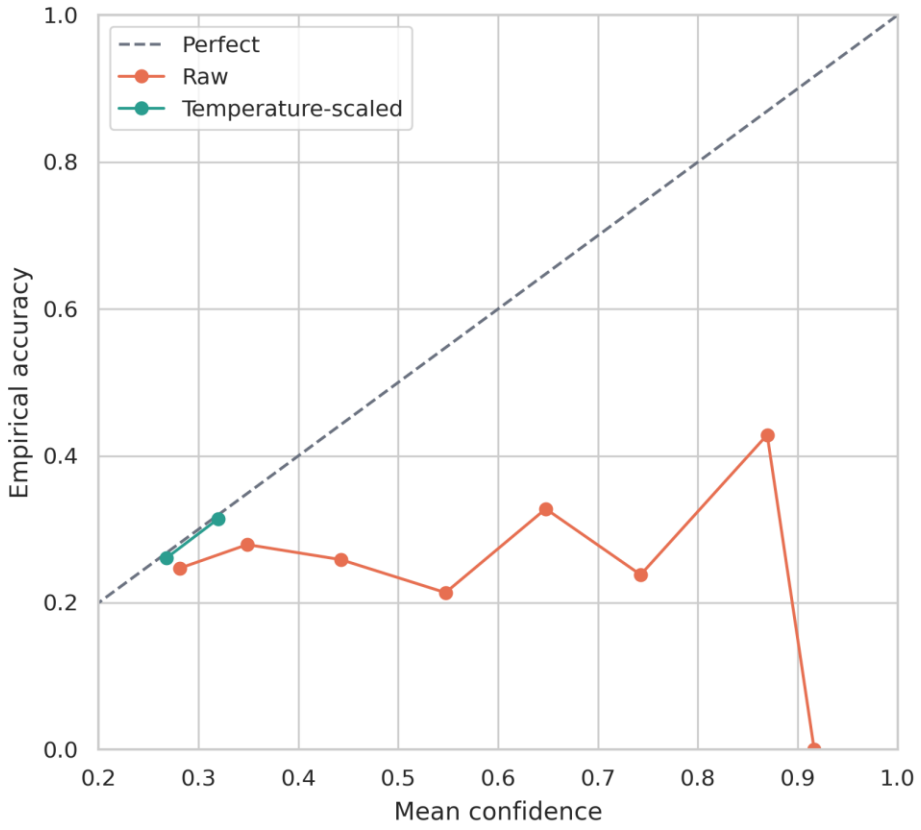

Figure 5. Reliability curves before and after temperature scaling for BM25-AllChoices Top-5.

These probability improvements did not create useful selective ordering. For the top-five system, standard AURC was 0.73456 and citation-aware AURC was 0.72778. Its full-coverage accuracy was 0.26382, but accuracy among the top 80%, 60%, and 40% by citation-aware selection score was 0.26806, 0.27095, and 0.26151. Decreasing coverage therefore failed to produce a consistent reduction in risk. BM25 top 1 behaved similarly. The gold-passage condition was different: accuracy rose from 0.49581 at full coverage to 0.55707 at 80%, 0.59497 at 60%, and 0.61506 at 40%. Its standard AURC was 0.39257 and its citation-aware AURC was 0.39462. The product score became informative when evidence was actually relevant.

Table 9. Selective prediction at fixed coverage and calibration-derived targets

System	Selection rule	Coverage	Selective risk	Selective accuracy
No Evidence	Top 100% product score	100.000%	75.628%	24.372%
No Evidence	Top 80% product score	79.983%	75.497%	24.503%
No Evidence	Top 60% product score	59.966%	76.536%	23.464%
No Evidence	Top 40% product score	40.034%	77.197%	22.803%
No Evidence	Calibration target 10%	0.084%	100.000%	0.000%
No Evidence	Calibration target 20%	0.084%	100.000%	0.000%
Gold Passage	Top 100% product score	100.000%	50.419%	49.581%
Gold Passage	Top 80% product score	79.983%	44.293%	55.707%
Gold Passage	Top 60% product score	59.966%	40.503%	59.497%
Gold Passage	Top 40% product score	40.034%	38.494%	61.506%
Gold Passage	Calibration target 10%	1.843%	31.818%	68.182%
Gold Passage	Calibration target 20%	6.868%	36.585%	63.415%
BM25-AllChoices Top-5	Top 100% product score	100.000%	73.618%	26.382%
BM25-AllChoices Top-5	Top 80% product score	79.983%	73.194%	26.806%
BM25-AllChoices Top-5	Top 60% product score	59.966%	72.905%	27.095%
BM25-AllChoices Top-5	Top 40% product score	40.034%	73.849%	26.151%
BM25-AllChoices Top-5	Calibration target 10%	0.084%	100.000%	0.000%
BM25-AllChoices Top-5	Calibration target 20%	0.084%	100.000%	0.000%

Note. Fixed-coverage rows order items by the product of calibrated maximum probability and the chosen answer's support score. Calibration-target thresholds use the same product and are chosen separately within each fold before application to its test fold.

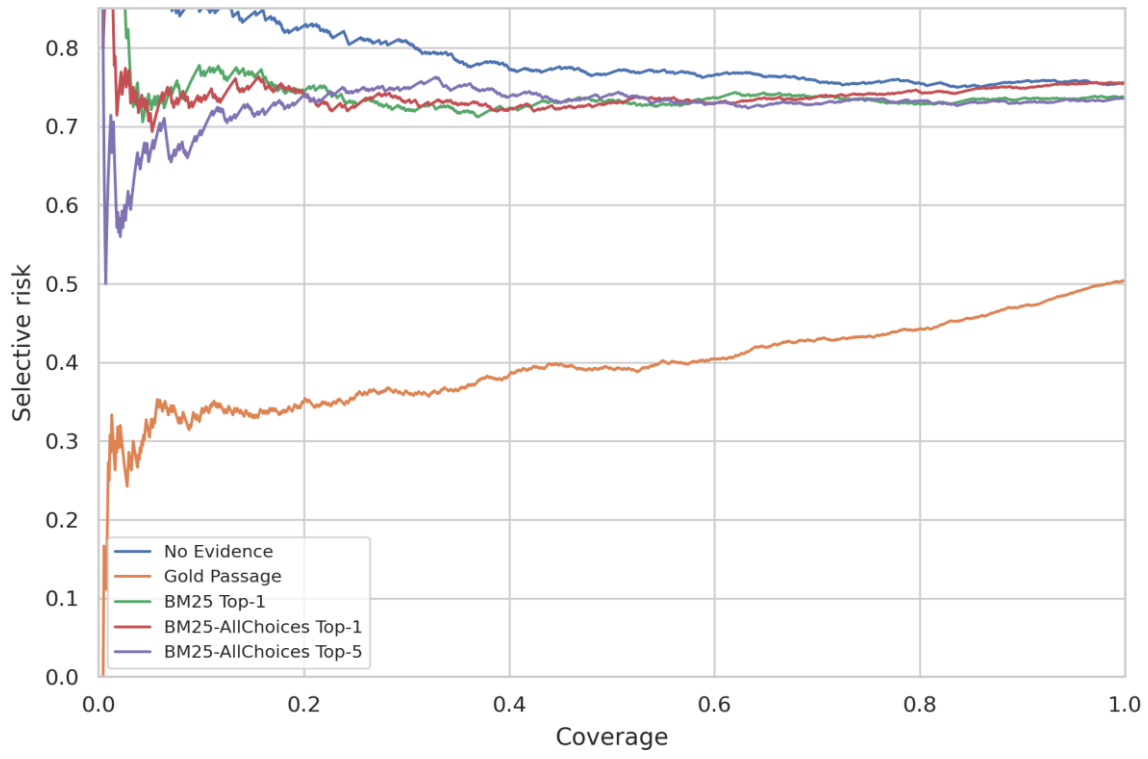

Figure 6. Citation-aware risk–coverage curves for no evidence, gold evidence, and retrieved-evidence conditions.

Calibration-fold product-score thresholds were also unstable. At a nominal 10% risk, the top-five system answered one item (coverage 0.00084) and that answer was wrong; the same occurred at the 20% target. The gold-passage system answered 1.843% of items under the 10% target but realized 31.818% risk, and answered 6.868% under the 20% target but realized 36.585% risk. These outcomes demonstrate why a low aggregate ECE cannot be read as a selective-risk guarantee (Si et al., 2022). They also agree with findings that confidence-based abstention can degrade under distributional change (Kamath et al., 2020; Varshney et al., 2022). In this experiment, abstention should not be deployed as a safety gate because the citation-aware selection score did not reliably rank retrieved-evidence answers by correctness.

The distinction between calibration and policy validation is consistent with counterfactual and off-policy evaluation research: a score can be statistically usable while the action rule selected from it remains unsafe or inefficient (Mu et al., 2025; Ye et al., 2026). Privacy-robust uplift estimation and identifier-loss modeling show that the evaluation design must also survive missing linkage information (Bai, Wang, et al., 2026; Bai & Wu, 2026). Evidence-grounded admission policies, strategy-controlled emotional-support responses, rubric-grounded educational scoring, and student-retention rationales all make the operating policy an explicit object of evaluation (Xin, 2026a, 2026b; Y. Zhang & Zhou, 2026; Zhao et al., 2026). The failed legal-QA thresholds therefore remain substantive results rather than being hidden by the improved ECE.

4.6 Evidence citation, support compatibility, and error decomposition

Citation correctness remained the most restrictive bottleneck. BM25 top 1 cited the strict gold passage on 0.251% of complete questions. AllChoices top 1 reached 0.754%, and AllChoices top 5 reached 3.183%. Equivalent-text rates were identical. The gold-passage condition was 100% by construction, while no evidence was 0%. These measures require annotated support rather than a source that merely looks legal.

Table 10. Citation validity and support-compatibility diagnostics

Panel A. Citation outcomes

System	Depth	Strict gold-ID rate	Equivalent-text rate	Mean support proxy
Gold Passage	1	100.000%	100.000%	0.6182
No Evidence	0	0.000%	0.000%	0.4483
BM25 Top-1	1	0.251%	0.251%	0.5135
BM25-AllChoices Top-1	1	0.754%	0.754%	0.6772
BM25-AllChoices Top-5	5	3.183%	3.183%	0.6834

Panel B. Held-out support-proxy validation

Metric	Value
Held-out choice–passage pairs	4,776
Positive pairs	1,194
ROC–AUC	0.6690
Average precision	0.4189
F1 at threshold 0.5	0.4578

Note. Citation rates test whether the supplied context contains the annotated gold identifier or a normalized-text equivalent. The support proxy is a diagnostic compatibility model, not a legal-entailment judgment.

The mean support-compatibility proxy was 0.5135 for BM25 top 1, 0.6772 for AllChoices top 1, 0.6834 for top 5, 0.6182 for gold evidence, and 0.4483 without evidence. The fact that retrieved conditions exceeded gold evidence reveals a limitation of the proxy: word overlap and negation agreement can assign high scores to a persuasive distractor. On 4,776 held-out gold-evidence choice pairs, the proxy achieved ROC-AUC 0.6690, average precision 0.4189, and F1 0.4578 at 0.5. It provides a moderately informative diagnostic, not citation entailment. Claim-level evaluation methods such as those studied by T. Gao et al. (2023), Honovich et al. (2022), and Min et al. (2023) motivate stronger future checks.

Evidence-chain and claim-verification research motivates retaining source identity, aligning claims to supporting spans, and exposing unsupported transitions (Jin, 2025a; C. Li et al., 2024; W. Su, Chen, et al., 2026; Zhong, Chen, et al., 2025). Retrieval-quality prediction can warn before answer generation but cannot prove support (R. Zhang et al., 2025), matching the observed proxy behavior.

Error decomposition makes the pipeline failure explicit. For BM25 top 1, 878 of 1,194 items (73.534%) combined a missed gold passage with a wrong answer; only one item (0.084%) combined retrieved gold evidence with a correct answer. For AllChoices top 1, corresponding counts were 895 (74.958%) and 3 (0.251%). Top-five evidence improved the favorable cell to 16 items (1.340%) and reduced missed-gold/wrong-answer cases to 857 (71.776%), but 22 items retrieved gold evidence and still received a wrong answer. The 299 missed-gold/correct cases show that some answers arose from question and choice cues rather than annotated evidence.

Table 11. Retrieval–answer error decomposition

System	Pipeline outcome	Count	Proportion
BM25 Top-1	Gold retrieved + answer correct	1	0.084%
BM25 Top-1	Gold retrieved + answer wrong	2	0.168%
BM25 Top-1	Gold missed + answer correct	313	26.214%
BM25 Top-1	Gold missed + answer wrong	878	73.534%
BM25-AllChoices Top-1	Gold retrieved + answer correct	3	0.251%
BM25-AllChoices Top-1	Gold retrieved + answer wrong	6	0.503%
BM25-AllChoices Top-1	Gold missed + answer correct	290	24.288%
BM25-AllChoices Top-1	Gold missed + answer wrong	895	74.958%
BM25-AllChoices Top-5	Gold retrieved + answer correct	16	1.340%
BM25-AllChoices Top-5	Gold retrieved + answer wrong	22	1.843%
BM25-AllChoices Top-5	Gold missed + answer correct	299	25.042%
BM25-AllChoices Top-5	Gold missed + answer wrong	857	71.776%

Note. Each system's four rows sum to N = 1,194. Gold retrieval is evaluated at the evidence depth supplied to the answer selector.

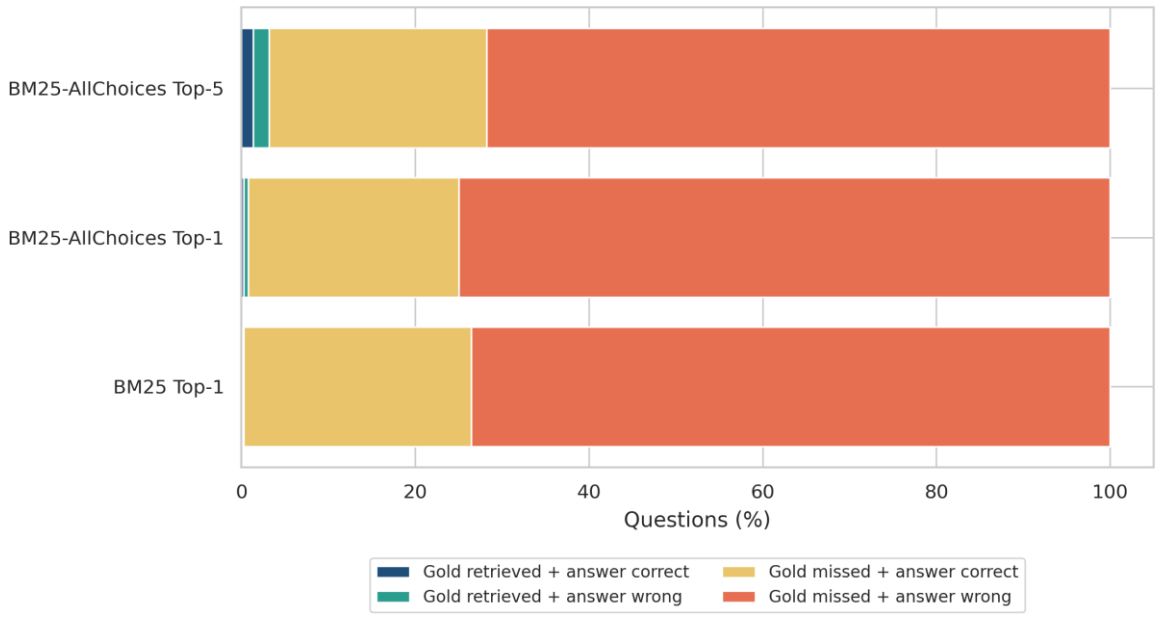

Figure 7. End-to-end error decomposition for BM25 and answer-choice-expanded evidence conditions.

4.7 Computational profile, limitations, and implications

The BM25 matrix required 352.25 seconds and 36.20 million nonzero weights. LSI required 160.03 seconds and a 329,024,640-byte matrix. Retrieval and evaluation took 575.81 seconds: 111.48 for BM25, 172.33 for choice BM25, 95.21 for LSI, 59.68 for feedback, 124.04 for rule-space expansion, 9.59 for evaluation, and 0.42 for encoding.

Table 12. Measured computational profile

Panel A. Index and representation builds

Artifact	Measured content	Seconds	Stored size
Corpus database	Physical-row parse, identifiers, hashes, and FTS5	64.31	973.70 MiB
BM25 matrix	36,195,706 nonzero weights	352.25	183.54 MiB
LSI embeddings	856,835 × 96 float32	160.03	313.78 MiB

Panel B. Full retrieval experiment

Stage	Seconds
Encode base and expanded queries	0.42
BM25 base retrieval	111.48
BM25 all-choice retrieval	172.33
LSI retrieval conditions	95.21
Pseudo-relevance feedback	59.68
Rule-space expansion	124.04
Evaluation and saved outputs	9.59
Retrieval experiment total	575.81

Note. Wall-clock measurements characterize this execution and include the explicit total saved by the retrieval program; index-build times are reported separately.

Several limitations bound the conclusions. The benchmark measures U.S. multistate bar-exam reasoning, not client counseling, jurisdiction-specific legal advice, or open-ended drafting. The single annotated gold passage is a defensible reference point but not necessarily the only legally useful authority. Exact text-hash equivalence addresses duplicate rows but cannot recognize a different passage stating the same rule. Passage truncation at 4,000 characters can remove relevant material from exceptionally long rows. The 96-dimensional LSI model is deliberately lightweight and should not be treated as representative of neural retrievers such as DPR or E5. The rule-space mapping is supervised by gold passages from training groups, so its result is evidence about cross-fitted representation transfer, not zero-shot retrieval.

The answer selector is also intentionally limited. It compares fixed answer choices through lexical and structural features and does not generate a legal analysis. Its gold-evidence accuracy of 49.581% shows both that evidence matters and that matching evidence to an outcome requires reasoning beyond term compatibility. The support score is not a validated legal entailment model. Finally, fold-specific risk thresholds were estimated from only about 237–240 calibration items per fold, making extreme low-risk thresholds unstable. These constraints do not weaken the central diagnostic finding: retrieval misses dominate the end-to-end error, and calibration cannot compensate for absent or distracting evidence.

Future evaluation should preserve this decomposition while testing stronger neural retrievers, claim-level legal entailment, adversarial retrieval, and privacy-aware evidence handling. The interface should also be evaluated as part of the safety case: evidence cards, risk dashboards, and editable rationales can help a reviewer inspect support, but they must not convert an uncertain model output into apparent authority (Jin, 2025b; Z. Li et al., 2025; W. Su, Rao, et al., 2026; Wu et al., 2025). This direction joins technical reliability to a verifiable human-review path.

5. Conclusions

This full-corpus study evaluated 1,195 Bar Exam QA retrieval questions, 1,194 complete answer items, and all 856,835 passages with a transparent retrieval and answer-selection pipeline. The measured overlap confirmed the task's central difficulty: mean query–gold TF–IDF cosine similarity was only 0.05095, and 182 pairs had zero cosine overlap. Under the fixed 40,000-term BM25 index, plain question retrieval reached 1.339% Recall@10. Appending all answer choices produced the strongest retrieval result at 5.105%, while cross-fitted rule-space expansion raised base BM25 to 1.841%. Classical 96-dimensional LSI did not supply a competitive semantic alternative, and equal-weight fusion with it reduced base BM25 performance.

Downstream results exposed a much larger evidence gap. Gold evidence raised answer accuracy to 49.581%, but the best retrieved-evidence system reached only 26.382%, a nonsignificant 2.010-percentage-point improvement over no evidence.

Temperature scaling sharply reduced ECE, yet confidence did not rank retrieved-evidence answers well enough for abstention: retaining only the top 40% reduced rather than improved top-five accuracy. Citation results were stricter still, with gold or exact-equivalent evidence present in only 3.183% of top-five contexts. The dominant error—71.776% of items—was a missed gold passage followed by a wrong answer.

The practical conclusion is that legal RAG evaluation should keep retrieval, answer correctness, citation validity, calibration, and selective risk separate. A fluent or confident answer cannot repair missing authority, and a low ECE does not demonstrate a safe abstention policy. The strongest direction suggested by this run is improved rule-aware retrieval: richer neural encoders, carefully validated query reformulation, and legal entailment reranking can be tested against the same full corpus, while preserving leakage-group separation and exact citation checks. Multi-regulation RAG and trust-calibrated multilingual retrieval offer complementary governance and access perspectives, but they should be evaluated on legal-domain evidence and abstention criteria before deployment (Y. Chen & Xu, 2026; J. Li & Zhou, 2026). Until retrieval recall and evidence-conditioned reasoning improve together, calibrated confidence should be treated as a diagnostic signal rather than a basis for autonomous legal answering.

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. *Information, 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] Chalkidis, I., Fergadiotis, M., Manginas, N., Katakalou, E., & Malakasiotis, P. (2021). Regulatory compliance through Doc2Doc information retrieval: A case study in EU/UK legislation where text similarity has limitations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3498–3511). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.305

[5] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon Reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[6] 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

[7] Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. *Journal of Technology Informatics and Engineering, 4*(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490

[8] 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

[9] 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

[10] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/JACS.2024.40407

[11] 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

[12] Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 758–759). Association for Computing Machinery. https://doi.org/10.1145/1571941.1572114

[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, L., Ma, X., Lin, J., & Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1762–1777). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.99

[16] 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

[17] 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. *Advances in Neural Information Processing Systems, 36*, 44123–44279. https://proceedings.neurips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., & Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3905–3920). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.287

[24] Jagerman, R., Zhuang, H., Qin, Z., Wang, X., & Bendersky, M. (2023). Query expansion by prompting large language models. *arXiv*. https://doi.org/10.48550/arXiv.2305.03653

[25] Jia, P., Liu, Y., Zhao, X., Li, X., Hao, C., Wang, S., & Yin, D. (2024). MILL: Mutual verification with large language models for zero-shot query expansion. 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. 2498–2518). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.138

[26] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics, 9*, 962–977. https://doi.org/10.1162/tacl_a_00407

[27] Jin, J. (2025a). 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

[28] Jin, J. (2025b). 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

[29] Jin, J., Huang, T., & Lu, S. (2024a). 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

[30] Jin, J., Huang, T., & Lu, S. (2024b). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[31] 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

[32] Karpukhin, V., Oğuz, 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

[33] 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

[34] Lei, Y., Cao, Y., Zhou, T., Shen, T., & Yates, A. (2024). Corpus-steered query expansion with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 393–401). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-short.34

[35] 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). Curran Associates. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[36] 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

[37] 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*. Advance online publication. https://doi.org/10.51903/jtie.v5i2.538

[38] 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

[39] 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

[40] 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

[41] 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

[42] Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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

[48] Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2024). Hallucination-free? Assessing the reliability of leading AI legal research tools. *arXiv*. https://doi.org/10.48550/arXiv.2405.20362

[49] 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

[50] Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

[51] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., 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

[52] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 4*(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

[53] 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

[54] 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

[55] 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

[56] Robertson, S. E., & Spärck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27*(3), 129–146. https://doi.org/10.1002/asi.4630270302

[57] Rosa, G. M., Rodrigues, R. C., Lotufo, R., & Nogueira, R. (2021). Yes, BM25 is a strong baseline for legal case retrieval. *arXiv*. https://doi.org/10.48550/arXiv.2105.05686

[58] 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

[59] Si, C., Zhao, C., Min, S., & Boyd-Graber, J. (2022). Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 2814–2829). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.204

[60] Su, H., Yen, H., Xia, M., Shi, W., Muennighoff, N., Wang, H.-y., Liu, H., Shi, Q., Siegel, Z. S., Tang, M., Sun, R., Yoon, J., Arik, S. O., Chen, D., & Yu, T. (2025). BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/7a0f8055c838df8e62329a76c7c6403d-Abstract-Conference.html

[61] 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

[62] 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

[63] 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

[64] 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

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

[66] 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

[67] Varshney, N., Mishra, S., & Baral, C. (2022). Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 1995–2002). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.158

[68] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv*. https://doi.org/10.48550/arXiv.2212.03533

[69] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 11897–11916). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.642

[70] Wang, L., Yang, N., & Wei, F. (2023). Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9414–9423). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.585

[71] Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. *International Journal of Graphic Design, 3*(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

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

[73] 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

[74] 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

[75] 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

[76] 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

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

[78] Xin, Q. (2026d). 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

[79] Xin, Q. (2026e). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI Online Retail. *J-INTECH (Journal of Information and Technology), 14*(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

[80] Xin, Q. (2026f). 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

[81] 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

[82] Ye, T., Chang, X., & Zhong, E. (2025). Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST$_2$24. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3701

[83] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (Small). *Journal of Technology Informatics and Engineering, 5*(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

[84] 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

[85] 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

[86] 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

[87] 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

[88] Zhang, Y., & Zhang, H. (2025a). 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

[89] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

[90] 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

[91] 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

[92] 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*. Advance online publication. https://doi.org/10.51903/jtie.v5i2.545

[93] 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

[94] 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

[95] 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

[96] 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

[97] 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

[98] 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

[99] 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), iaae020. https://doi.org/10.1093/imaiai/iaae020

[100] 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

[101] 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). https://proceedings.mlr.press/v258/zhong25a.html

[102] 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

[103] Zhou, B., Jin, J., & Zhao, D. (2025). 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

[104] 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

[105] 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