Rule of Law Studies Journal

Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces

Jing Li[1], Ashley Zhou[2]
[1] School of Law, New York University, NY, USA
[2] Data Science, Columbia University, NY, USA

Received: 18 March 2026 | Revised 20 May 2026 | Accepted: 14 June 2026

ABSTRACT

Background: The growing use of artificial intelligence (AI) in cross-border digital commerce has intensified compliance challenges arising from overlapping regulatory frameworks, particularly the European Union Artificial Intelligence Act (EU AI Act), the General Data Protection Regulation (GDPR), and the Digital Operational Resilience Act (DORA). Although Retrieval-Augmented Generation (RAG) systems are increasingly adopted to support legal and compliance decision-making, retrieval failures may undermine legal certainty and regulatory accountability.
Aim: This study evaluates the reliability of legal RAG systems in identifying and ranking regulatory obligations required for AI product counsel and examines retrieval as a legal governance mechanism rather than merely a technical process.
Method: An empirical audit was conducted using 139 English ComplianceBench scenarios and a corpus of 289 article- and annex-level units extracted from official EU AI Act, GDPR, and DORA texts. Seven transparent lexical retrieval methods were assessed using Hit@k, Recall@10, Mean Reciprocal Rank (MRR), nDCG@10, and citation-sufficiency metrics. External validation was performed using 300 AIReg-Bench technical-documentation excerpts.
Result: The hybrid lexical retriever achieved the strongest overall performance (Hit@1 = 0.360, Hit@10 = 0.770, Recall@10 = 0.392, MRR = 0.493, nDCG@10 = 0.344). While at least one relevant obligation appeared within the top ten results for 77.0% of queries, only 39.2% of the complete labelled obligation set was recovered on average. Validation on AIReg-Bench showed substantially higher retrieval effectiveness, with TF-IDF word retrieval achieving Hit@10 = 1.000 and MRR = 0.974.
Conclusion: Retrieval quality is a critical legal-governance control point that directly affects the reliability of AI-assisted compliance advice. Evidence-first architectures, citation-sufficiency thresholds, confidence-sensitive escalation, and human review mechanisms are necessary to support accountable AI product counsel in cross-border digital commerce.

Keywords: AI Product Counsel; Retrieval-Augmented Generation; Regulatory Compliance.

Cite this article: 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.

INTRODUCTION

The rapid integration of artificial intelligence (AI) into digital commerce has transformed how organizations manage regulatory compliance, legal risk assessment, and product governance across jurisdictions. AI-enabled systems increasingly support decisions concerning market entry, product classification, technical documentation, consumer protection, and operational compliance (Kalathil, 2025; Kumar & Suthar, 2024). As AI technologies become embedded within financial services, e-commerce platforms, and automated decision-making environments, organizations must navigate an increasingly complex regulatory landscape (Pazouki et al., 2025; Rane et al., 2024). Recent regulatory developments, particularly the adoption of the European Union Artificial Intelligence Act (EU AI Act), have significantly expanded legal obligations associated with AI deployment. At the same time, organizations remain subject to established frameworks such as the General Data Protection Regulation (GDPR) and sector-specific regulations including the Digital Operational Resilience Act (DORA). These overlapping legal regimes create substantial challenges for legal practitioners and compliance professionals seeking to ensure regulatory conformity. Consequently, the development of reliable AI-assisted legal support systems has become an increasingly important issue for cross-border digital commerce.

Despite remarkable advances in large language models, the use of AI in legal and compliance contexts remains controversial. AI systems frequently generate responses that appear legally persuasive while lacking adequate factual or regulatory support (Ashley, 2017; Clarke, 2019). Such failures are particularly problematic in legal environments because incorrect advice may influence regulatory decisions, internal governance processes, risk-management strategies, and organizational accountability. Unlike ordinary information retrieval applications, legal decision-making requires traceable reasoning supported by authoritative legal sources. Several scholars have therefore raised concerns regarding the reliability, transparency, and explainability of AI-generated legal outputs (Citron, 2008; Citron & Pasquale, 2014; Dancy & Zalnieriute, 2026; Goktas, 2024). The challenge becomes even greater when compliance assessments involve multiple regulatory regimes operating simultaneously. As AI-generated legal assistance becomes increasingly integrated into organizational workflows, ensuring the legitimacy of underlying evidence has become a critical governance concern.

Retrieval-Augmented Generation (RAG) has emerged as one of the most widely adopted approaches for reducing hallucination and improving the factual grounding of large language models. By combining language generation with external knowledge retrieval, RAG systems seek to provide responses supported by verifiable evidence rather than relying solely on model parameters (Abo El-Enen et al., 2025; Huang, & Huang, 2026; Lewis et al., 2020). In legal applications, retrieval mechanisms offer an additional layer of transparency because users can inspect the sources underlying generated recommendations. However, the effectiveness of this approach depends fundamentally on retrieval quality. When controlling legal authorities are omitted, poorly ranked, or overshadowed by irrelevant materials, the resulting answer may still be misleading despite containing citations. In such circumstances, the retrieval layer effectively determines which legal obligations are available for reasoning and which are excluded before human review occurs. Consequently, retrieval should be understood not merely as a technical process but as a governance mechanism that directly affects legal reliability and accountability.

The significance of retrieval quality becomes particularly evident in contemporary regulatory environments characterized by overlapping legal obligations. The EU AI Act introduces a risk-based

framework governing AI systems and imposes requirements related to risk management, technical documentation, transparency, human oversight, and post-market monitoring (European Parliament & Council, 2024). Simultaneously, the GDPR establishes obligations concerning personal-data processing, automated decision-making, transparency, and data-protection impact assessments (European Parliament & Council, 2016). DORA further introduces operational resilience and ICT-risk management obligations for financial institutions (European Parliament & Council, 2022). In practice, many AI-enabled products—including credit-scoring systems, insurance technologies, biometric applications, and digital onboarding platforms—fall within the scope of multiple regulatory frameworks simultaneously. Compliance therefore depends not only on identifying a single relevant provision but on accurately mapping interconnected obligations across several legal regimes. Failure to retrieve these obligations may undermine both regulatory compliance and organizational decision-making.

The governance implications of retrieval failures can be better understood through the lens of rule-of-law theory. Classical scholars have emphasized that legal systems must provide public, stable, predictable, and administrable rules that enable individuals and organizations to understand their rights and obligations (Dicey, 1885; Fuller, 1964; Raz, 1977). Contemporary interpretations further stress transparency, justification, accountability, and reviewability as essential components of lawful governance (Bingham, 2011; Waldron, 2011). These principles are increasingly relevant in AI-supported legal environments where automated systems influence compliance assessments and regulatory interpretations. Recent AI governance frameworks similarly emphasize explainability, accountability, fairness, and human oversight as core requirements for trustworthy AI systems (Barocas et al., 2023; NIST, 2023; OECD, 2024). From this perspective, the adequacy of retrieval is not simply a matter of system performance but a prerequisite for maintaining legal certainty and institutional accountability.

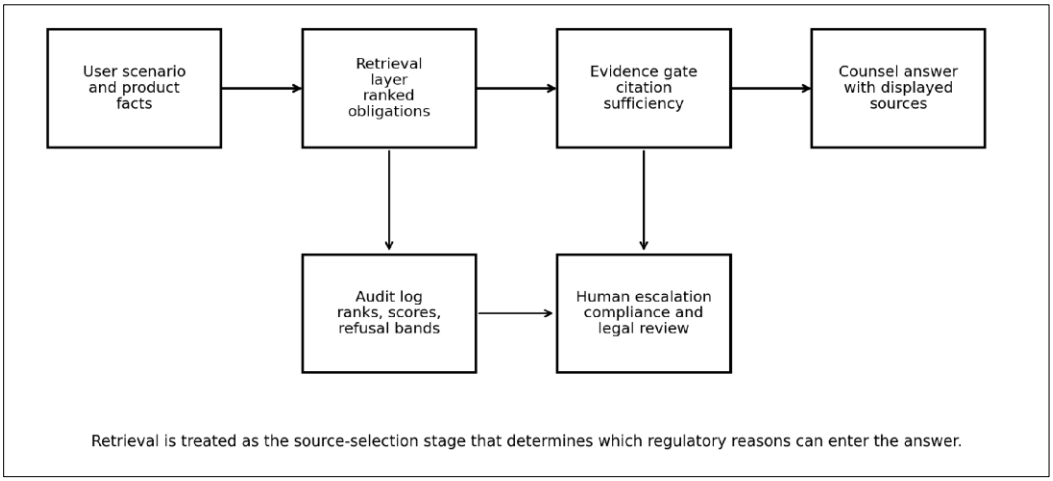

Figure 1. Evidence-first governance framework for AI product counsel RAG

To operationalize these principles, this study adopts an evidence-first governance perspective that positions retrieval as a legal accountability mechanism within AI-assisted compliance systems. Figure 1 presents the proposed Evidence-First Governance Framework. The framework conceptualizes retrieval as the process through which product facts and compliance questions are transformed into a ranked set of regulatory evidence. Retrieved sources are subsequently evaluated

through an evidence gate that determines whether the available legal authorities are sufficient to support answer generation, whether the output should be treated as unverified, or whether escalation to human legal review is necessary. Audit logs preserve rankings, confidence scores, thresholds, and citation histories to ensure transparency and post hoc reviewability. By embedding retrieval decisions within a structured governance workflow, the framework translates rule-of-law principles into practical design requirements for AI product counsel systems.

Existing scholarship has extensively explored legal information retrieval, legal analytics, and retrieval-enhanced AI systems. Foundational research has demonstrated the effectiveness of probabilistic retrieval approaches such as BM25 and related ranking mechanisms for identifying relevant legal materials (Manning et al., 2008; Robertson & Zaragoza, 2009). More recent studies have examined retrieval-augmented generation in legal and regulatory contexts, highlighting improvements in citation accuracy and answer grounding (Ashley, 2017; Lewis et al., 2020). The emergence of benchmark datasets such as ComplianceBench and AIReg-Bench has further enabled systematic evaluation of legal retrieval performance within compliance-oriented environments (Augustyniak, 2026; Marino et al., 2025). Although these studies have contributed substantially to methodological development, the dominant focus remains on retrieval effectiveness and answer accuracy. Comparatively less attention has been devoted to the governance implications of retrieval failures and citation insufficiency.

Several important limitations therefore remain unresolved. Existing studies rarely evaluate retrieval systems using rule-of-law criteria such as legal certainty, auditability, source sufficiency, and reviewability. Furthermore, most evaluations prioritize whether at least one relevant source is retrieved rather than whether the complete set of applicable obligations is identified. The governance consequences of omitted authorities, ranking errors, and incomplete obligation mapping remain insufficiently understood, particularly in multi-regulatory environments where compliance depends on integrating obligations across different legal frameworks. As a result, current literature provides limited insight into how retrieval failures affect the reliability of AI-assisted legal compliance systems. This gap is especially significant for AI product counsel applications operating within complex cross-border regulatory environments.

Accordingly, this study investigates how reliably legal RAG systems retrieve and rank regulatory obligations required for AI product counsel across the EU AI Act, GDPR, and DORA. Using ComplianceBench and AIReg-Bench datasets, the study evaluates multiple retrieval approaches and examines retrieval performance through a rule-of-law governance perspective. The research contributes theoretically by extending discussions on AI governance and legal informatics beyond traditional retrieval metrics toward questions of legal accountability and citation sufficiency. Practically, it offers evidence-based recommendations for evidence-first architectures, citation-sufficiency thresholds, confidence-sensitive escalation mechanisms, ranked citation interfaces, and human-review procedures. Through these contributions, the study seeks to support the development of more transparent, auditable, and legally accountable AI systems for cross-border digital commerce.

METHOD

Datasets and regulatory corpus

The main experiment uses ComplianceBench, a legal compliance benchmark containing 266 scenarios across EU AI Act, GDPR, DORA, and related regulatory questions. The main evaluation is

limited to the English scenarios with at least one in-scope gold obligation in the three-law corpus. This produces 139 evaluated scenarios. The Polish scenarios are used only as a robustness diagnostic because the regulatory corpus was built from English official texts.

The retrieval corpus consists of official EU AI Act, GDPR, and DORA texts. Each regulation was extracted into article-level units; AI Act annexes were also included because high-risk classification and technical-documentation duties frequently turn on annex references. Recitals were not scored as controlling obligations because the audit focuses on operative articles and annexes that product counsel would cite in an obligation map. Table 1 reports the measured dataset profile, and Table 2 reports the distribution of labelled in-scope obligations by law and unit type. Figure 2 visualizes the corpus composition.

Table 1. Dataset and regulatory corpus profile.

Dataset property	Measured value
Corpus units	289
AI Act corpus units	126
GDPR corpus units	99
DORA corpus units	64
ComplianceBench total scenarios	266
ComplianceBench English scenarios	140
English scenarios with at least one in-scope gold obligation	139
Mean in-scope gold obligations per English scenario	4.6
Median corpus unit tokens	288
Mean corpus unit tokens	401.8
95th percentile corpus unit tokens	1117.6
Mean query tokens	72.6
AIReg-Bench documentation excerpts	300

Table 2. Labelled in-scope obligations by law and unit type

Law	Unit	Gold obligation mentions	Unique obligations
AI Act	Annex	137	4
AI Act	Article	430	41
DORA	Article	16	8
GDPR	Article	51	13

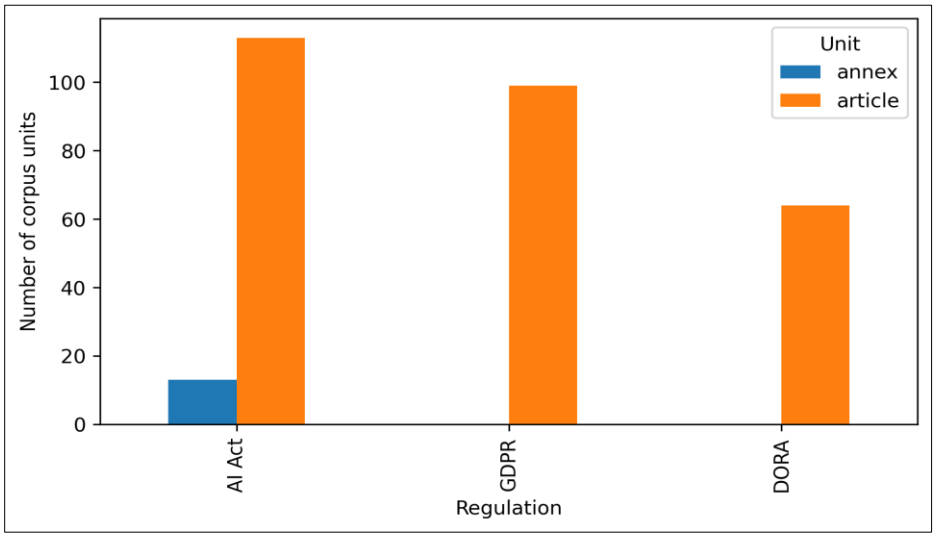

Figure 2. Corpus units by regulation and unit type.

AIReg-Bench is used as an external product-documentation validation set. It contains 300 technical-documentation excerpts organized by the EU AI Act article to which the excerpt pertains. The validation task asks whether the retrieval layer can recover the target AI Act article from the documentation text. To avoid direct label leakage, the article heading at the beginning of each documentation excerpt was removed before retrieval.

A methodological qualification is necessary. ComplianceBench labels are treated as benchmark evidence for evaluation, not as a final legal ontology. Several scenarios contain more than one legally relevant provision, and some background law may be unlabelled or outside the three-law corpus. For that reason, the article uses the phrase labelled in-scope obligation rather than claiming that every unlabelled article is legally irrelevant. This responds to the same problem that legal reviewers often identify in single-gold-passage legal RAG: a miss under the benchmark may understate or overstate useful retrieval depending on whether alternative sources would support a correct answer. The results should therefore be interpreted as an audit of citation placement against the benchmark's identified obligations, not as a complete assessment of all possible legal support.

Retrieval systems

The audit compared seven reproducible lexical retrieval systems. The systems were chosen because they are transparent, stable, and inspectable. The hybrid weights were fixed before evaluation to emphasize BM25 title+text while testing whether word, character, and title-overlap views add complementary signal. Table 3 lists the exact configurations.

Table 3. Retriever configurations used in the empirical audit

System	Reproducible configuration
Lexical overlap	Binary word-count vectorizer over title-weighted regulatory unit text; score is shared-token count.
TF-IDF word	Word unigram/bigram TF-IDF, sublinear term frequency, L2 cosine similarity, 60,000-feature cap.
TF-IDF char	Character 3-5 gram TF-IDF with word-boundary padding, L2 cosine similarity, 20,000-feature cap.
BM25 text only	Sparse Okapi BM25 over regulatory unit text; k1 = 1.5, b = 0.75, 60,000-feature cap.
BM25 title+text	Same BM25 model with the regulation, article/annex number, and heading prepended once to the text.
Hybrid lexical	0.45 BM25 title+text + 0.25 TF-IDF word + 0.25 TF-IDF char + 0.05 title-term overlap after row-wise min-max scaling.
RRF lexical ensemble	Reciprocal rank fusion of TF-IDF word, TF-IDF char, and BM25 title+text with k = 60.

Evaluation protocol and metrics

Each query was formed from the ComplianceBench scenario text plus the question. Gold reasoning was used only to identify benchmark obligations and was not used to construct retrieval queries. This prevents answer-label leakage. Each query was evaluated against all 289 regulatory units. The retrieval pipeline used Python 3.13 with NumPy 2.3.5, pandas 2.2.3, scikit-learn 1.8.0, and matplotlib 3.10.8. Bootstrap confidence intervals used 2,000 resamples over questions with seed 42.

Because the task is multi-label, the metrics distinguish between finding at least one controlling obligation and finding the whole obligation set. Hit@10 asks whether any labelled obligation appears in the ten-unit context window. Recall@10 asks what share of the labelled obligation set appears there. Complete@10 asks whether the full labelled set appears there. Table 4 defines the metrics and their legal interpretation.

Table 4. Evaluation metrics and legal interpretation

Metric	Operational meaning in this audit
Hit@k	Whether at least one labelled in-scope regulatory obligation appears within the first k retrieved units.
Recall@10	Mean share of labelled in-scope obligations retrieved in the top ten.
Complete@10	Whether the complete labelled in-scope obligation set appears in the top ten.
MRR	Mean reciprocal rank of the first labelled obligation; rewards early placement of a controlling source.
nDCG@10	Rank-discounted gain for multi-label binary relevance within the top ten.
Top-10 miss rate	Share of questions for which no labelled in-scope obligation appears in the ten-unit context window.

RESULT AND DISCUSSION

Results

Overall retrieval performance

The overall results show that lexical retrieval can often surface at least one relevant obligation, but it remains weak as a complete obligation-mapping mechanism. The strongest system by MRR was the hybrid lexical retriever. It achieved Hit@1 = 0.360, Hit@10 = 0.770, Recall@10 = 0.392, MRR = 0.493, and nDCG@10 = 0.344. BM25 title+text reached a slightly higher Hit@10 of 0.784, but with lower Hit@1 and MRR. Lexical overlap achieved high Hit@10 but weak first-source placement, which makes it a poor default for citation-first legal advice. Table 5 reports the main metrics, while Figures 3 and 4 show Hit@k and MRR.

The difference between Hit@10 and Recall@10 is legally important. A product counsel assistant may retrieve some relevant obligation while still omitting most of the obligation set required for a reliable cross-regulation answer. In a single-regulation question, partial retrieval may be useful. In a multi-regulation question, partial retrieval can be misleading if the answer appears comprehensive while omitting GDPR, DORA, or AI Act duties that matter to the product decision.

Table 5. Overall retrieval and citation-placement metrics on ComplianceBench English scenarios

Method	Hit@1	Hit@10	Recall@10	MRR	nDCG@10	Median first-gold rank	Top-10 miss rate
Hybrid lexical	0.360	0.770	0.392	0.493	0.344	3	0.230
RRF lexical ensemble	0.317	0.727	0.364	0.458	0.317	3	0.273
BM25 title+text	0.295	0.784	0.393	0.443	0.327	4	0.216
BM25 text only	0.266	0.770	0.388	0.417	0.313	4	0.230
TF-IDF word	0.273	0.612	0.281	0.382	0.245	5	0.388
TF-IDF char	0.209	0.662	0.325	0.358	0.262	5	0.338
Lexical overlap	0.122	0.827	0.414	0.310	0.272	5	0.173

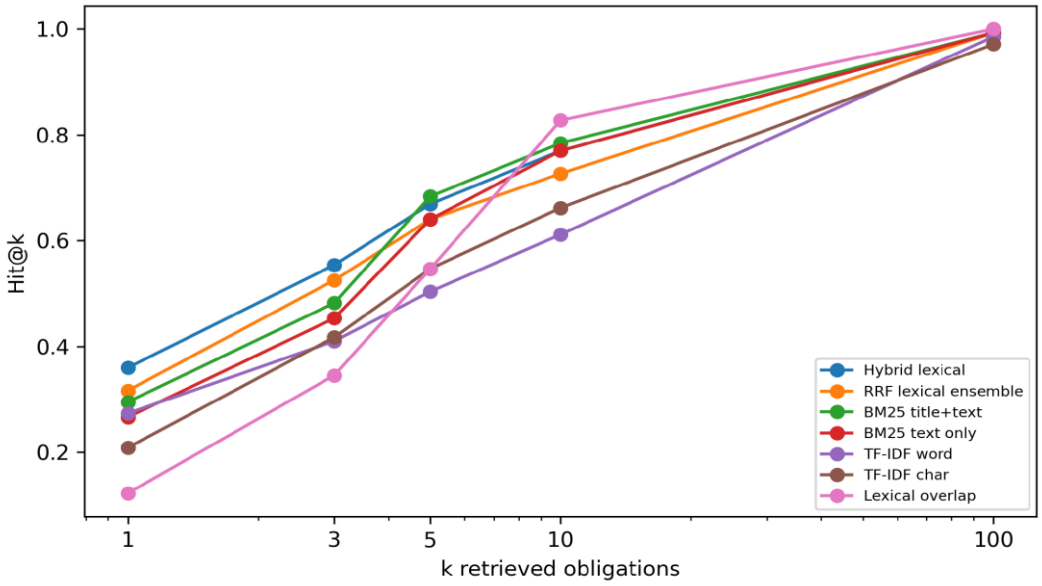

Figure 3. Hit@k curves across retrieval systems

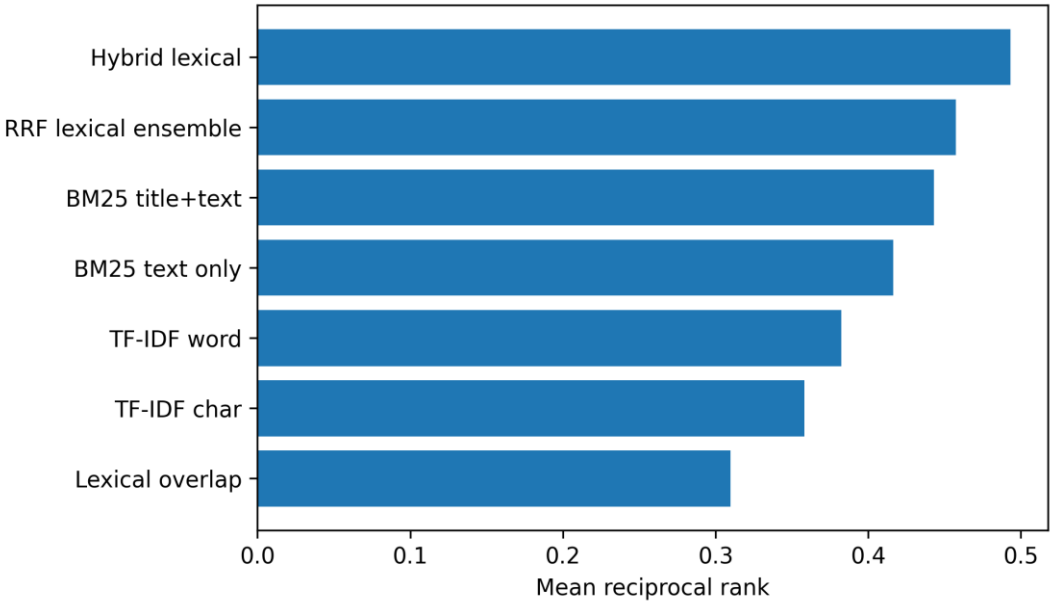

Figure 4. Mean reciprocal rank across retrieval systems

Bootstrap intervals confirm that the measured differences are not driven by one or two questions. Table 6 reports intervals for the primary legal-design metrics. Even under the strongest MRR system, the top-ten context missed every labelled in-scope obligation in 23.0% of questions and recovered less than half of the labelled obligation set on average.

Table 6. Bootstrap confidence intervals for primary retrieval metrics

Method	Hit@1	Hit@10	MRR	Recall@10
Hybrid lexical	0.360 [0.281, 0.439]	0.770 [0.698, 0.835]	0.493 [0.425, 0.561]	0.392 [0.340, 0.444]
RRF lexical ensemble	0.317 [0.237, 0.396]	0.727 [0.655, 0.799]	0.458 [0.395, 0.526]	0.364 [0.313, 0.420]
BM25 title+text	0.295 [0.216, 0.374]	0.784 [0.712, 0.849]	0.443 [0.383, 0.506]	0.393 [0.341, 0.446]

Method	Hit@1	Hit@10	MRR	Recall@10
BM25 text only	0.266 [0.194, 0.338]	0.770 [0.705, 0.842]	0.417 [0.353, 0.481]	0.388 [0.333, 0.439]
TF-IDF word	0.273 [0.201, 0.353]	0.612 [0.532, 0.698]	0.382 [0.319, 0.449]	0.281 [0.232, 0.332]
TF-IDF char	0.209 [0.144, 0.273]	0.662 [0.583, 0.741]	0.358 [0.299, 0.422]	0.325 [0.273, 0.376]
Lexical overlap	0.122 [0.072, 0.180]	0.827 [0.763, 0.885]	0.310 [0.264, 0.358]	0.414 [0.358, 0.471]

Failure decomposition

The failure decomposition translates ranks into system-design consequences. A top-1 success means the first displayed citation is a labelled obligation. A rank-order failure means at least one labelled obligation appears within ranks 2-10, but the lead source is not controlling. Candidate-set weakness means the first labelled obligation appears only in ranks 11-100. Retrieval collapse means no labelled obligation appears in the top 100. Table 7 and Figures 5-6 show this decomposition.

Under the hybrid retriever, 50 questions were top-1 successes, 57 had the first labelled obligation in ranks 2-10, 31 had the first labelled obligation in ranks 11-100, and one question had no labelled obligation in the top 100. These categories matter because they call for different interventions. Rank-order failures may be addressed through reranking or citation-interface design. Candidate-set weakness may require a larger candidate pool, query rewriting, or lawyer-facing search support. Retrieval collapse is a refusal or escalation problem.

Table 7. Retrieval failure decomposition across systems

Method	Top-1 correct	Ranks 2-10	Ranks 11-100	Not in top 100	Top-10 miss rate
Hybrid lexical	50	57	31	1	0.230
RRF lexical ensemble	44	57	37	1	0.273
BM25 title+text	41	68	29	1	0.216
BM25 text only	37	70	31	1	0.230
TF-IDF word	38	47	52	2	0.388
TF-IDF char	29	63	43	4	0.338
Lexical overlap	17	98	24	0	0.173

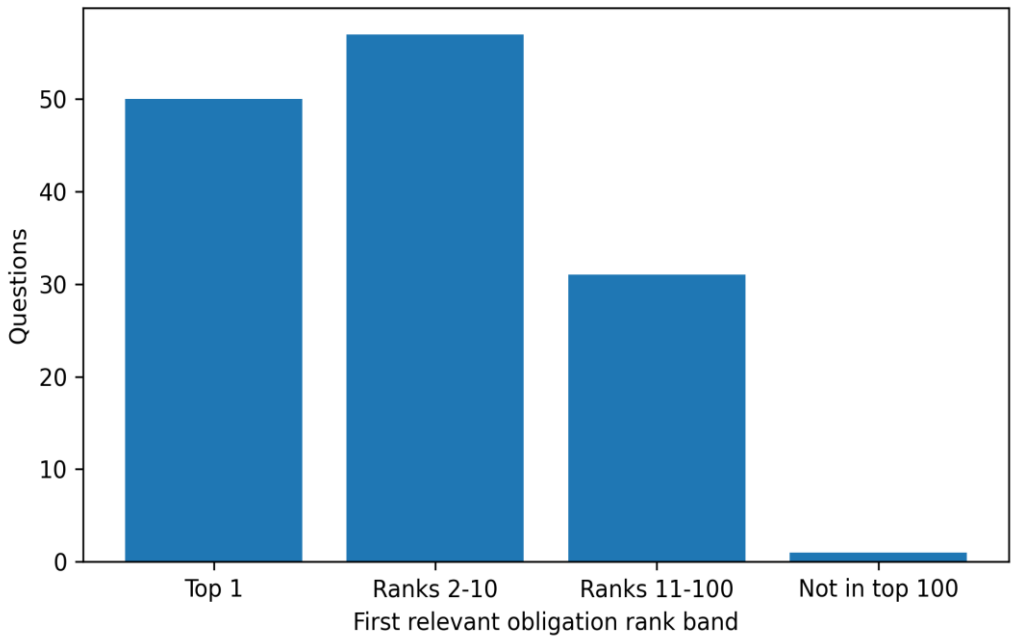

Figure 5. Rank distribution of labelled obligations for the strongest MRR system

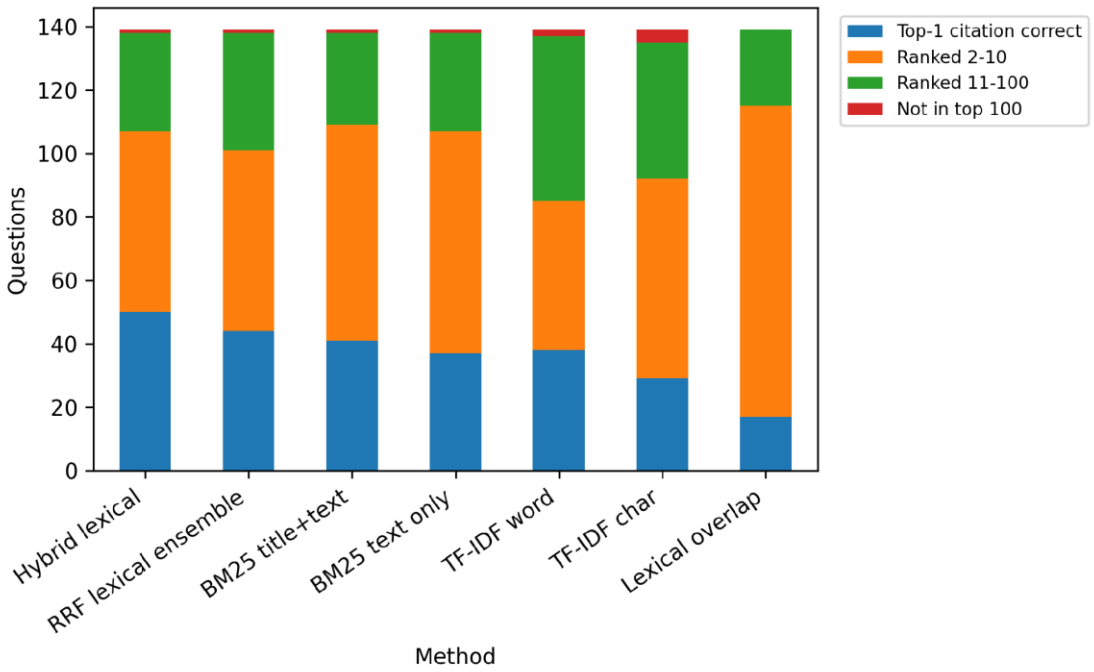

Figure 6. Error decomposition across retrieval systems

Lexical mismatch, task diagnostics, and language robustness

Lexical overlap remains a decisive stressor. In the lowest-overlap quartile, the hybrid retriever reached Hit@10 = 0.514 and MRR = 0.217. In the highest-overlap quartile, Hit@10 rose to 1.000 and MRR to 0.727. Table 8 and Figure 7 show the pattern. This result is practically important because product teams often describe systems in business language, while legal obligations use regulatory terms such as deployer, high-risk system, post-market monitoring, or ICT third-party risk.

Task-level diagnostics are useful, but they should not be overread as stable doctrinal conclusions. Some task groups are small. Table 9 therefore reports them as design diagnostics rather than final category-level claims. The most consistent pattern is that direct prohibition, provider-obligation, and citation-verification tasks are easier than incident-reporting, financial-services overlap, or cross-regulatory documentation tasks. Table 10 shows the bilingual robustness check. Performance drops on Polish scenarios because the query language no longer matches the English corpus, which supports the need for multilingual regulatory corpora or translation-aware retrieval in cross-border deployment.

Table 8. Best-system performance by query-obligation lexical overlap quartile

Overlap quartile	n	Hit@1	Hit@10	MRR	Recall@10	Median rank
Q1 lowest overlap	35	0.086	0.514	0.217	0.315	10
Q2	35	0.314	0.657	0.436	0.364	4
Q3	34	0.441	0.912	0.595	0.527	2
Q4 highest overlap	35	0.600	1	0.727	0.365	1

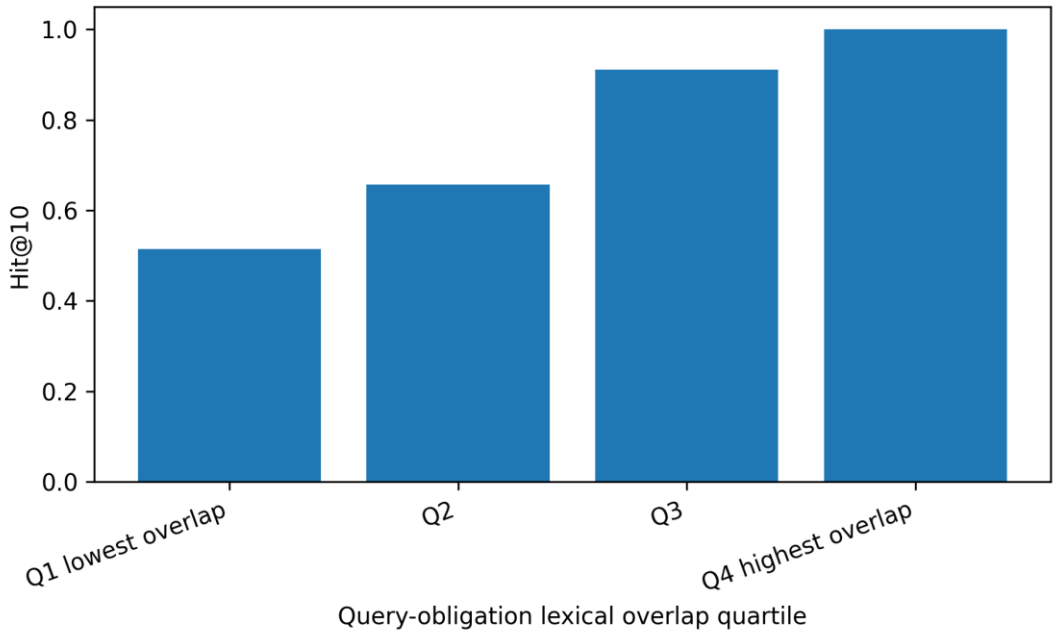

Figure 7. Hit@10 by lexical-overlap quartile for the strongest MRR system

Table 9. Task-level diagnostics for the strongest MRR system

Task	n	Hit@1	Hit@10	MRR	Recall@10	Complete@10	Median rank
A2	35	0.229	0.600	0.334	0.329	0.057	7
A1	19	0.579	1	0.715	0.719	0.421	1
D1	16	0.625	1	0.757	0.445	0	1
B1	12	0.667	1	0.774	0.284	0	1
C1	8	0.125	0.750	0.288	0.241	0.125	6.500
F1	8	0.250	0.625	0.383	0.372	0.250	4.500
E1	8	0.250	0.500	0.405	0.312	0.125	7
F3	7	0.143	0.429	0.216	0.262	0.143	61
C2	7	0	0.571	0.139	0.252	0.143	7
B2	5	0.800	1	0.867	0.578	0.200	1
F2	5	0	0.600	0.219	0.367	0.200	7
F4	5	0.400	1	0.667	0.440	0.200	2
E2	4	0.250	1	0.500	0.199	0	3

Table 10. Robustness check by query language

Language	n	Hit@1	Hit@10	MRR	Recall@10	Median rank
en	139	0.360	0.770	0.493	0.392	3
pl	97	0.155	0.474	0.271	0.245	11

Citation burden and user-interface implications

Citation display creates a second governance problem. Showing more citations increases the chance that at least one labelled obligation is present, but it also increases the burden placed on the user. Table 11 and Figure 8 show the trade-off. Under the hybrid retriever, a ten-citation display included at least one labelled obligation for 77.0% of questions, but it recovered only 39.2% of the labelled obligation set on average and added 8.57 non-gold citations per answer. A long list of

citations can therefore look accountable while shifting legal research back onto the lawyer or compliance officer.

The interface should not treat citations as decorative proof. Citations should be ranked by confidence, grouped by legal issue or regulation, and labelled where possible as controlling, supporting, background, or unverified. For cross-regulation advice, the interface should show whether each identified regime has at least one cited obligation. If the answer says that both the AI Act and GDPR apply, the citation pane should not contain only AI Act sources.

Table 11. Citation sufficiency and non-gold citation burden

Method	k cited	Any gold included	Mean gold recall	Complete gold set	Mean non-gold citations	All-citation failure
Hybrid lexical	1	0.360	0.127	0.043	0.640	0.640
Hybrid lexical	5	0.669	0.289	0.094	4.058	0.331
Hybrid lexical	10	0.770	0.392	0.137	8.568	0.230
BM25 title+text	1	0.295	0.114	0.050	0.705	0.705
BM25 title+text	5	0.683	0.293	0.101	4.079	0.317
BM25 title+text	10	0.784	0.393	0.137	8.568	0.216
BM25 text only	1	0.266	0.106	0.050	0.734	0.734
BM25 text only	5	0.640	0.278	0.101	4.137	0.360
BM25 text only	10	0.770	0.388	0.129	8.590	0.230

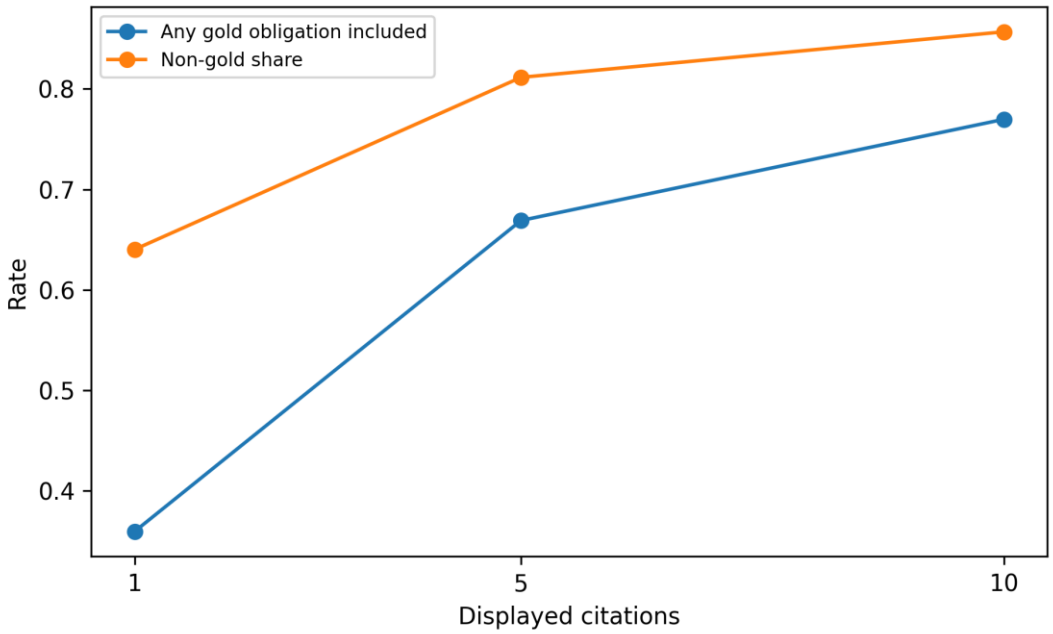

Figure 8. Citation sufficiency and non-gold citation trade-off for the strongest MRR system

External validation on AIReg-Bench technical documentation

The AIReg-Bench validation asks a narrower product-documentation question: given a technical-documentation excerpt, can the retrieval layer recover the EU AI Act article that the documentation concerns? The task is easier than the ComplianceBench obligation-mapping task because each excerpt is organized around a specific article. Nonetheless, it is useful because it connects the framework to a concrete AI product counsel workflow: reviewing technical documentation for risk management, data governance, logging, human oversight, accuracy, robustness, and cybersecurity.

Table 12 reports the validation metrics and Table 13 reports performance by target article. TF-IDF word retrieval reached Hit@1 = 0.950 and Hit@10 = 1.000. The strong result does not eliminate the main compliance-risk finding; it shows that retrieval can perform well when product documentation is article-aligned. The contrast with ComplianceBench indicates that the hardest task is not always retrieving from documentation that mirrors a regulation. The harder task is mapping messy product facts and multi-regulation questions onto a complete obligation set. Figure 9 shows the article-level validation pattern.

Table 12. External validation metrics on AIReg-Bench technical-documentation excerpts

Method	Hit@1	Hit@10	MRR	nDCG@10	Top-10 miss rate
TF-IDF word	0.950	1	0.974	0.981	0
Hybrid lexical	0.813	1	0.892	0.919	0
RRF lexical ensemble	0.770	1	0.855	0.891	0
TF-IDF char	0.703	0.980	0.801	0.844	0.020
BM25 title+text	0.713	0.977	0.794	0.837	0.023
BM25 text only	0.713	0.973	0.793	0.835	0.027
Lexical overlap	0	0.603	0.223	0.302	0.397

Table 13. AIReg-Bench validation by target EU AI Act article

Gold article	n	Hit@1	Hit@10	MRR	Median rank
AI_ACT_ART_10	60	1	1	1	1
AI_ACT_ART_12	60	1	1	1	1
AI_ACT_ART_14	60	0.750	1	0.869	1
AI_ACT_ART_15	60	1	1	1	1
AI_ACT_ART_9	60	1	1	1	1

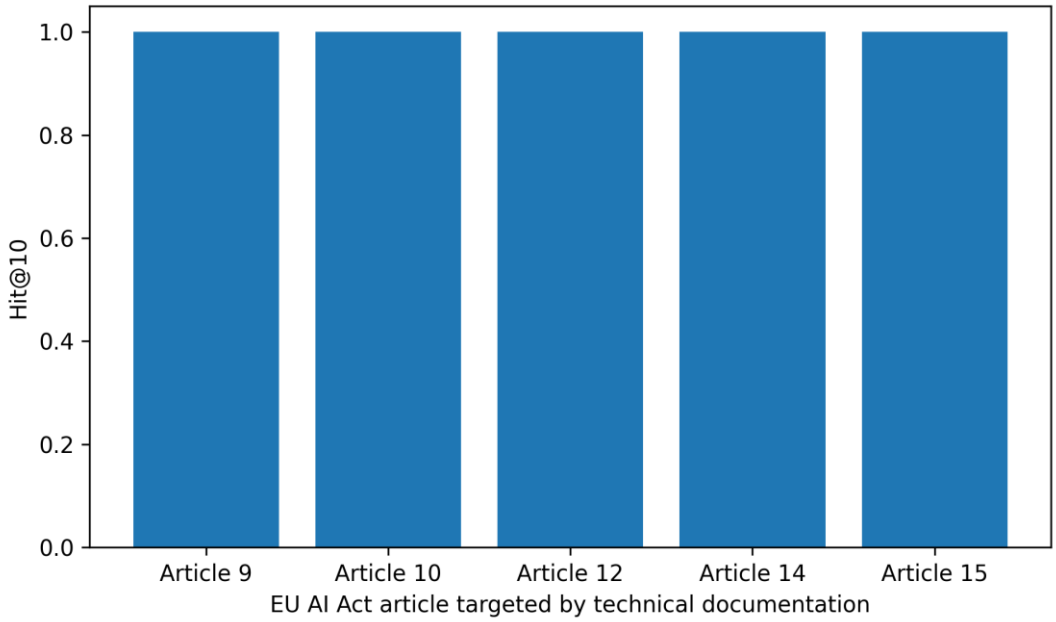

Figure 9. AIReg-Bench Hit@10 by target EU AI Act article

Design implications for AI product counsel

The results support an evidence gate rather than a generic disclaimer. The gate should assess whether the retrieved set contains a plausible controlling source, whether sources are ranked high enough to be used by the generator, whether each asserted regulatory regime has citation support, and whether source scores are consistent rather than flat or noisy. A practical gate can combine rank thresholds, retrieval-score margins, topic consistency, citation agreement between retriever and answerer, and coverage of the expected obligation set. The threshold need not be universal. A platform merchant-support tool, an internal counsel assistant, and a regulatory filing workflow can set different sufficiency bands, but each should disclose the band and record the decision.

The gate must also avoid unnecessary refusals. A system should not refuse merely because the benchmark label is absent if another legally sufficient source is present. In deployment, that issue can be handled by human verification, issue-level grouping, reranking, and source-equivalence checks. In research, it requires richer multi-source labels and legal expert review. The present audit therefore treats a missing labelled obligation as evidence of risk, not as proof that no legal answer could be supported.

Human review should focus on high-risk failure bands. Low-overlap scenarios, incident-reporting tasks, multi-regulation financial-services questions, and multilingual queries show elevated risk. A product counsel system should surface those cases for legal review before producing definitive advice. The audit log should preserve query text, retrieved units, scores, ranks, displayed citations, generation inputs, refusals, and human overrides. Those records are not merely debugging material; they are the compliance trail that allows a firm to explain how the system supported a legal decision.

Discussion

The findings demonstrate that retrieval performance should be evaluated not only as a technical measure of information access but also as a determinant of legal reliability in AI-assisted compliance systems. Although the hybrid lexical retriever achieved the strongest overall ranking performance, its ability to recover the complete set of labelled obligations remained limited. This indicates that retrieval success cannot be assessed solely by whether at least one relevant provision is identified. From a legal-governance perspective, the omission of applicable obligations may affect the completeness of compliance assessments and the quality of subsequent legal reasoning. This interpretation is consistent with rule-of-law scholarship, which emphasizes legal certainty, predictability, and accountability as essential conditions for legitimate decision-making (Fuller, 1964; Raz, 1977; Waldron, 2011). The findings therefore support the study's central proposition that retrieval quality functions as a governance mechanism because it shapes the legal evidence available for both automated reasoning and human review.

A particularly important finding is the discrepancy between obligation identification and obligation completeness. While the strongest retrieval systems frequently placed at least one relevant obligation within the top-ranked results, they recovered substantially fewer obligations when evaluated against the complete labelled obligation set. This suggests that retrieval systems may provide a partial representation of the applicable legal framework, especially in environments where compliance depends on multiple overlapping regulations. Such a pattern is highly relevant to contemporary AI governance because organizations increasingly operate under interconnected regimes, including the EU AI Act, GDPR, and DORA. The result extends previous work on Retrieval-

Augmented Generation by demonstrating that citation presence alone does not necessarily guarantee comprehensive legal support (Buffa et al., 2025; Lewis et al., 2020). Consequently, legal reliability should be assessed not only through retrieval accuracy but also through the sufficiency of retrieved obligations to support compliance-oriented decision-making.

The analysis of retrieval failures provides further insight into the operational risks of legal RAG systems. The distinction between rank-order failures, candidate-set weaknesses, and retrieval collapse indicates that retrieval errors emerge from different underlying mechanisms and therefore require different governance responses. Cases in which relevant obligations appear but are ranked too low may be addressed through improved ranking strategies or citation prioritization. In contrast, failures arising from weak candidate generation suggest limitations in the retrieval stage itself. This finding aligns with broader discussions of algorithmic accountability, which emphasize that transparency alone is insufficient unless systems incorporate mechanisms capable of identifying and mitigating operational risks (Citron & Pasquale, 2014; NIST, 2023). The results therefore imply that governance controls should be embedded within retrieval workflows rather than applied only after answer generation. Such an approach strengthens the evidentiary foundation upon which legal recommendations are constructed.

Another significant finding concerns the influence of lexical overlap between regulatory language and user queries. Retrieval effectiveness improved substantially when the terminology used in compliance scenarios closely matched the language of regulatory provisions. Conversely, performance declined in low-overlap conditions, where business-oriented descriptions differed from formal legal terminology. This outcome is consistent with established information-retrieval theory, which highlights the importance of lexical correspondence between queries and documents (Manning et al., 2008; Robertson & Zaragoza, 2009). However, within legal-compliance contexts, lexical mismatch has broader implications because it may prevent legally relevant obligations from entering the evidentiary set available for analysis. The result suggests that compliance-oriented retrieval systems face challenges beyond conventional search tasks, particularly when regulatory concepts are expressed through diverse operational vocabularies. Accordingly, future legal AI systems may benefit from mechanisms that improve semantic alignment between business descriptions and legal requirements.

The multilingual robustness analysis further highlights the contextual nature of legal retrieval performance. The observed decline in effectiveness when Polish-language scenarios were evaluated against an English-language corpus suggests that language mismatch remains a substantial challenge for cross-border compliance applications. This finding is particularly important because regulatory compliance frequently operates across multilingual jurisdictions where legal obligations must be interpreted and applied in different linguistic contexts. Existing AI governance frameworks emphasize reliability and fairness across deployment environments (Barocas et al., 2023; Leon, 2026; OECD, 2024), and the present findings support those concerns by showing that retrieval performance may vary considerably when language conditions change. Consequently, evaluations conducted solely within monolingual environments may overestimate the practical robustness of legal AI systems. Future research should therefore investigate multilingual retrieval architectures and translation-aware compliance workflows capable of supporting broader regulatory ecosystems.

The citation-sufficiency analysis offers an additional perspective on transparency and accountability in legal AI systems. Increasing the number of displayed citations improved the

likelihood that at least one relevant obligation would be included; however, it also increased the number of non-relevant citations presented to users. This finding suggests that transparency involves a trade-off between evidentiary coverage and user burden. Simply providing more citations does not necessarily improve legal reliability if users must independently determine which sources are controlling and which are peripheral (Lins & Sunyaev, 2023; Wong & Floridi, 2023). Similar concerns have been raised in studies of automated decision-making, where procedural transparency may not translate directly into meaningful accountability (Citron, 2008; Citron & Pasquale, 2014). The present results therefore indicate that citation design should prioritize relevance, ranking quality, and evidentiary sufficiency rather than citation quantity alone. In legal-compliance environments, effective transparency depends on the quality of evidence presented rather than the volume of references displayed.

The external validation results provide a useful contrast to the primary compliance benchmark. Retrieval performance was substantially stronger when evaluated using technical-documentation excerpts that were closely aligned with specific regulatory articles. This suggests that lexical retrieval methods remain effective when product documentation explicitly reflects regulatory structures and terminology. Nevertheless, the difference between the validation results and the primary compliance benchmark demonstrates that article-level retrieval is considerably easier than mapping real-world product facts to a complete set of regulatory obligations. This distinction is important because compliance practice rarely involves perfectly structured documentation. Instead, legal practitioners must interpret complex operational contexts that frequently involve overlapping obligations and incomplete information. The findings therefore reinforce the argument that benchmark success should not automatically be interpreted as evidence of comprehensive legal-compliance capability.

Taken together, these findings contribute to the growing literature on AI governance, legal informatics, and retrieval-augmented systems by highlighting the central role of retrieval in shaping legal accountability. Previous studies have demonstrated the value of retrieval mechanisms for improving factual grounding and reducing unsupported generation (Ashley, 2017; Lewis et al., 2020; Surden, 2014). The present study extends this literature by showing that retrieval quality influences not only answer accuracy but also the completeness of the legal evidence available for compliance assessment. Rather than viewing retrieval as a purely technical component, the results support its conceptualization as an evidentiary control layer within AI-assisted legal systems. Although the study is limited to a specific regulatory corpus and a set of predominantly lexical retrieval methods, it provides empirical evidence that accountable AI product counsel requires evidence-oriented retrieval, citation-sufficiency assessment, auditable decision pathways, and human oversight mechanisms. In this way, the study offers a practical and theoretical foundation for developing more transparent and legally reliable AI systems in cross-border digital commerce.

CONCLUSION

Multi-regulation RAG for AI product counsel should be evaluated as a legal governance system, not merely as an information-retrieval component. The retrieval layer determines which regulatory obligations are available for reasoning, citation, human review, and audit. When it omits or buries controlling obligations, the downstream answer may appear fluent while lacking the sources needed for regulatory certainty. The main Compliance Bench audit found that the best MRR system placed a

labelled in-scope obligation first for 36.0% of questions and placed at least one labelled obligation in the top ten for 77.0%. That is useful but insufficient for reliable obligation mapping: the same system recovered only 39.2% of the labelled obligation set on average in the top ten. The AIReg-Bench validation shows that article-aligned technical documentation can be much easier to retrieve, but product counsel questions remain harder when they require cross-regulation reasoning over messy product facts. Regulatory due diligence by design therefore requires evidence-first architecture. AI product counsel systems should expose retrieval confidence, record source rankings, distinguish controlling from background citations, require human review for low-confidence or multi-regulation outputs, and refuse or escalate when the evidence set is not sufficient. These safeguards translate rule-of-law values into concrete engineering controls for cross-border digital commerce.

AUTHOR CONTRIBUTION STATEMENT

Jing Li conceived the research project, developed the legal governance framework, interpreted the legal and regulatory implications, and wrote the manuscript.

Ashley Zhou developed and executed the retrieval pipeline, processed and analyzed the data, generated the figures and tables, and provided technical input on the empirical evaluation.

REFERENCES

Abo El-Enen, M., Saad, S., & Nazmy, T. (2025). A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Computing and Applications, 37*(33), 28191–28267. https://doi.org/10.1007/s00521-025-11666-9

Ashley, K. D. (2017). *Artificial intelligence and legal analytics: New tools for law practice in the digital age*. Cambridge University Press.

Augustyniak, L. (2026). *ComplianceBench* [Data set]. Hugging Face.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bingham, T. (2011). *The rule of law*. Penguin.

Buffa, M., Ferrara, A., Picascia, S., Riva, D., & Castano, S. (2025). Enhancing legal document building with retrieval-augmented generation. *Computer Law & Security Review, 59*, Article 106229. https://doi.org/10.1016/j.clsr.2025.106229

Citron, D. K. (2008). Technological due process. *Washington University Law Review, 85*(6), 1249–1313.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33.

Clarke, R. (2019). Regulatory alternatives for AI. *Computer Law & Security Review, 35*(4), 398–409. https://doi.org/10.1016/j.clsr.2019.04.008

Dancy, T., & Zalnieriute, M. (2026). AI and transparency in judicial decision making. *Oxford Journal of Legal Studies, 46*(1), 1–34. https://doi.org/10.1093/ojls/gqaf030

Dicey, A. V. (1885). *Introduction to the study of the law of the constitution*. Macmillan.

European Parliament and Council. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. *Official Journal of the European Union*.

European Parliament and Council. (2022). *Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (Digital Operational Resilience Act)*. *Official Journal of the European Union*.

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. *Official Journal of the European Union*.

Fuller, L. L. (1964). *The morality of law*. Yale University Press.

Goktas, P. (2024). Ethics, transparency, and explainability in generative AI decision-making systems: A comprehensive bibliometric study. *Journal of Decision Systems*, 1–29. https://doi.org/10.1080/12460125.2024.2410042

Huang, Y., & Huang, J. X. (2026). A survey on retrieval-augmented text generation for large language models. *ACM Computing Surveys, 58*(12), 1–38. https://doi.org/10.1145/3805774

Kalathil, L. (2025). AI-driven real-time compliance management in IoT-enabled retail operations. *International Journal of Emerging Trends in Computer Science and Information Technology*, 173–181.

Kumar, D., & Suthar, N. (2024). Ethical and legal challenges of AI in marketing: An exploration of solutions. *Journal of Information, Communication and Ethics in Society, 22*(1), 124–144. https://doi.org/10.1108/JICES-05-2023-0068

Leon, M. (2026). Lifecycle-based governance to build reliable ethical AI systems. *Systems Research and Behavioral Science*. Advance online publication. https://doi.org/10.1002/sres.70014

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.

Lins, S., & Sunyaev, A. (2023). Advancing the presentation of IS certifications: Theory-driven guidelines for designing peripheral cues to increase users' trust perceptions. *Behaviour & Information Technology, 42*(13), 2255–2278. https://doi.org/10.1080/0144929X.2022.2113432

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Marino, B., Hunter, R., Jamali, Z., Kalpakos, M. E., Kashyap, M., Hinton, I., Hanson, A., Nazir, M., Schnabl, C., Steffek, F., Wen, H., & Lane, N. D. (2025). *AIReg-Bench: Benchmarking language models that assess AI regulation compliance* (arXiv Preprint No. 2510.01474). arXiv.

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1)*. U.S. Department of Commerce.

OECD. (2024). *Recommendation of the Council on Artificial Intelligence*. OECD Legal Instruments (OECD/LEGAL/0449).

Pazouki, S., Jamshidi, M., Jalali, M., & Tafreshi, A. (2025). Artificial intelligence and digital technologies in finance: A comprehensive review. *Journal of Economics, Finance & Accounting Studies, 7*(2). https://doi.org/10.32996/jefas.2025.7.2.5

Rane, N., Choudhary, S. P., & Rane, J. (2024). Acceptance of artificial intelligence technologies in business management, finance, and e-commerce: Factors, challenges, and strategies. *Studies in Economics and Business Relations, 5*(2), 23–44. https://doi.org/10.48185/sebr.v5i2.1333

Raz, J. (1977). The rule of law and its virtue. *Law Quarterly Review, 93*, 195–211.

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

Surden, H. (2014). Machine learning and law. *Washington Law Review, 89*(1), 87–115.

Waldron, J. (2011). The rule of law and the importance of procedure. In J. E. Fleming (Ed.), *Getting to the rule of law* (pp. 3–31). New York University Press.

Wong, D., & Floridi, L. (2023). Meta's oversight board: A review and critical assessment. *Minds and Machines, 33*(2), 261–284. https://doi.org/10.1007/s11023-022-09613-x