International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DOCTRINE-CONSTRAINED INFORMATION RETRIEVAL FOR AI-ASSISTED LEGAL MARKET RESEARCH IN M&A DUE DILIGENCE

Raj Sonani

Cornell Tech, Cornell University

rs2397@cornell.edu

Abstract

When artificial intelligence tools are used for legal research in high-stakes transactional work—mergers and acquisitions (M&A) due diligence, in particular—they fail in ways that matter a great deal. Attorneys conducting due diligence must sift through large unstructured datasets to find material market information, and recent benchmarks show that even frontier large language models (LLMs) cannot reliably analyze regulatory filings. Industry surveys tell the same story: attorney trust in AI for liability-critical tasks remains low. We trace a root cause to system architecture. Today's approaches treat legal domain knowledge as training data or as post-generation evaluation criteria, but not as constraints woven into the information retrieval pipeline itself. We propose a framework that takes three established legal standards—the duty of competent representation (ABA Model Rule 1.1), the due diligence defense under Securities Act Section 11(b), and the materiality standard from TSC Industries, Inc. v. Northway, Inc.—and maps each one to a specific pipeline layer, constraining retrieval, filtering, and ranking before any content reaches the generation model. Out of this mapping come three failure classes for AI-assisted legal market research: factual misstatement, material omission, and what we term relevance indiscrimination. A prototype M&A research tool, built through a Cornell Tech industry collaboration with a law firm partner, was evaluated by attorney-annotators. Doctrine-level constraints at the filtering and ranking layers cut relevance indiscrimination by 31% relative to an unconstrained baseline, without degrading truthfulness or completeness. The upshot is that professional standards governing attorney conduct can do double duty as engineering specifications for the AI systems those attorneys rely on.

Keywords: *legal AI, information retrieval, regulatory compliance, materiality, domain-constrained systems, M&A due diligence, evaluation methodology*

1. INTRODUCTION

1.1 Context

Since the public release of large language models capable of generating and processing natural language at scale [43], the legal industry has moved quickly—and unevenly—toward adoption. Law firms, corporate legal departments, and regulatory compliance teams are all experimenting with AI-assisted tools for contract review, legal research, and document summarization [1, 2]. But a 2023 report from the Thomson Reuters Institute paints a more cautious picture: while awareness of generative AI was high, most firms had not yet woven AI into substantive legal workflows, held back by concerns about reliability, liability, and professional responsibility [3]. The gap between low-stakes applications (where mistakes are cheap to fix) and high-stakes

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

ones (where errors trigger professional liability, regulatory penalties, or direct financial harm) remains the field's central unsolved problem.

M&A due diligence sits at the demanding end of that spectrum. Attorneys working a deal must research the target company's market positioning, comparable transaction valuations, regulatory risk factors, and industry trends—drawing on market reports, financial filings, regulatory documents, analyst commentary, and industry publications to assemble assessments that guide transaction decisions worth billions of dollars [4]. Miss a material regulatory risk, and the client faces financial harm while the firm faces malpractice exposure [5].

The reliability gap is not hypothetical. In December 2023, Patronus AI published a benchmark showing that frontier LLMs—GPT-4 included—could not reliably analyze SEC filings [6]. The models produced factual errors, omitted material information, and generated plausible but unsupported claims. That finding landed hard in the legal industry, crystallizing what practitioners already suspected: AI systems that sound fluent and appear to comprehend legal text still fall short of the reliability bar for professional use where the cost of being wrong is severe.

1.2 Problem Identification

Look under the hood of current legal AI systems—whether general-purpose LLMs pressed into legal service or purpose-built tools like Harvey [7] and Thomson Reuters CoCounsel [8]—and they share one architectural assumption: the system's job is to process legal text and produce output, with quality control applied afterward through human review or automated checks [44]. The attorney reads the AI's work, catches mistakes, and fixes them. Generation first, evaluation second.

We challenge that assumption. In high-stakes legal work, generating unreliable output and hoping to catch it afterward is not just an efficiency drain—it is a liability trap. An attorney who relies on AI output that turns out to be materially incomplete faces professional consequences whether or not a review step was nominally in place. So the question we ask is not whether AI output can be evaluated after the fact, but whether the information retrieval pipeline can be built so that certain failure classes are structurally prevented—or at least flagged—before output ever reaches the attorney.

1.3 Core Observation

Legal doctrine already spells out, with considerable precision, what counts as adequate information delivery. These are not vague aspirations—they're standards refined through decades of case law, regulatory practice, and professional rule-making. The ABA's Model Rules define competent representation [9]. Section 11(b) of the Securities Act of 1933 specifies the due diligence defense for material omissions [10]. And the Supreme Court's holding in TSC Industries, Inc. v. Northway, Inc. sets the bar for what information is material in securities regulation [11]. All three are normative: they define what should be considered adequate, complete, and significant.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Yet none of these standards appear anywhere in the architecture of AI information retrieval systems that serve legal professionals. The gap is subtle but important. It is not simply that current legal AI fails to meet these standards—that much is an empirical fact. The deeper problem is that the standards are absent from the system design entirely. They live in the legal domain and have no counterpart in the engineering domain. What happens when you take those standards and embed them as architectural constraints inside the retrieval pipeline, rather than applying them as evaluation criteria to finished output? That is what we set out to test.

1.4 Contributions

Three contributions emerge from this work. First, a taxonomy of failure modes in AI-assisted legal market research, tied to established doctrines of professional liability. The taxonomy pins three failure classes—factual misstatement, material omission, and relevance indiscrimination—to specific layers of the retrieval pipeline. Second, a framework for turning doctrine-derived constraints into engineering specifications that operate at those layers, constraining retrieval, filtering, and ranking. Third, a prototype built through the Cornell Tech industry collaboration program, along with evaluation results showing the framework's effect on relevance indiscrimination in M&A due diligence research.

2. RELATED WORK

2.1 AI and NLP for Legal Applications

NLP research on legal applications has grown rapidly over the past decade. Early efforts centered on contract analysis and clause identification; the CUAD dataset [12] gave the field its first serious benchmark for legal contract understanding. Chalkidis et al. [13, 14] pushed legal text classification forward with transformer-based architectures [42], while Zheng et al. [15] showed that domain-specific pre-training on legal corpora lifts performance on downstream legal tasks.

With the arrival of large language models came a wave of legal task evaluations. Bommarito and Katz [16] tested GPT-3.5 and GPT-4 on bar examination questions; the models did well on multiple-choice items but stumbled on open-ended legal reasoning. Blair-Stanek et al. [17] found similar trouble with statutory reasoning—the precise, rule-bound logic that lawyers rely on daily. The LegalBench benchmark [18] attempted to bring order to this landscape with a comprehensive evaluation suite spanning multiple legal reasoning categories.

One pattern runs through all of this work: researchers optimize for task performance on legal text—accuracy, F1, task-specific metrics—but nobody stops to ask whether the system's output meets the professional adequacy standards that actually govern how attorneys can use it. The research question has been "did the AI get it right?" not "would this output satisfy the attorney's duty of competence?" Related questions, but not the same one, and the gap between them has real design consequences.

2.2 Information Retrieval in Domain-Specific Applications

Other fields have invested heavily in domain-specific information retrieval for high-stakes settings. Roberts et al. [19] showed that domain-adapted retrieval models beat general-purpose

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

systems for biomedical literature search. The TREC Clinical Decision Support track [20] built evaluation frameworks for medical IR where reliability is literally life-or-death. Shah et al. [21] tackled the distinct challenge of pulling structured financial data out of unstructured regulatory documents.

On the retrieval side, dense passage retrieval with learned embeddings (Karpukhin et al. [22]) now outperforms sparse methods like BM25 [23] on open-domain QA benchmarks. Lewis et al. [24] introduced retrieval-augmented generation (RAG) as a way to ground language model outputs in retrieved source documents—a pattern that has since become ubiquitous. Gao et al. [25] and Izacard and Grave [26] refined the approach with better retrieval strategies and fusion techniques.

Where domain-specific optimization has gone so far is mostly within the retrieval layer: fine-tuning embeddings on domain corpora, tweaking chunk sizes, training rerankers, expanding queries [27, 28]—all well-established IR techniques [41] that sharpen statistical alignment between queries and documents. What we haven't seen, anywhere in this literature, is a framework that treats normative domain standards—standards defining what information should be considered significant, as distinct from what is statistically likely to be relevant—as architectural constraints on the pipeline itself. Our contribution sits at the pipeline level, which makes it agnostic to what sits downstream. The framework applies whether the consumer of retrieved information is an LLM, a traditional search interface, or something else entirely.

2.3 Evaluation of AI Systems in Professional Contexts

How to evaluate AI systems deployed in professional settings is an active and unsettled question. The standard ML toolkit—accuracy, precision, recall, F1—tells you something, but not enough, when professional standards govern how output gets used [29]. In healthcare, Topol [30] and Rajkomar et al. [31] have argued that evaluation must account for clinical impact and the decision-making context surrounding the AI, not just its technical performance in isolation. The FDA's framework for Software as a Medical Device (SaMD) is one of the few cases where regulatory standards have been formally mapped to AI evaluation criteria [32].

Legal AI evaluation is further behind. LegalBench [18] offers task-level metrics, and Guha et al. [33] proposed quality metrics for legal text generation. But no existing framework turns professional standards—the competence duty, the due diligence standard, the materiality test—into evaluation metrics for AI systems. The question that matters for a legal AI tool is not "is this output accurate?" but "does this output meet the standard of competence, completeness, and materiality that the attorney's professional obligations demand?" We take that question seriously and propose a mapping as part of the framework.

3. FRAMEWORK

What follows is the core intellectual contribution: a framework for mapping legal doctrines onto information retrieval pipeline architecture. We start with a taxonomy of failure modes, move through the constrained pipeline design, and finish with a doctrine-derived evaluation rubric.

Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.1 Failure Taxonomy

Three distinct classes of failure recur in AI-assisted legal market research. Each one maps to a specific legal doctrine and localizes to a specific layer of the retrieval pipeline.

Failure Class 1: Factual Misstatement

The system states something that is flatly wrong—an incorrect comparable transaction valuation, a regulatory approval listed as complete when it is still pending, a market figure that does not match the source. Under ABA Model Rule 1.1, attorneys owe a duty of competent representation—and that includes making sure their tools don't inject errors they wouldn't have made on their own [9]. The failure originates at the generation layer, where the model produces content not grounded in retrieved sources. Researchers have studied this problem extensively under the headings of hallucination and factual grounding [34, 35].

Failure Class 2: Material Omission

The system fails to retrieve or surface information that is material to the transaction being analyzed. The corresponding doctrine is Securities Act Section 11(b), under which liability attaches for material omissions in registration statements and the due diligence defense requires demonstrating reasonable investigation [10]. We emphasize that Section 11(b) governs liability for specific statutory filings, not attorney work product generally; we invoke the standard here as a conceptual anchor for what constitutes adequate investigation rather than as a direct legal mapping. The underlying principle—that a professional who fails to conduct a reasonably thorough search cannot claim adequate diligence—is not unique to that statute: it is recognized independently in ABA Rule 1.1 for all legal representation and reflected in the broader common law duty of care—making it a robust normative anchor across legal contexts. In an M&A context, examples include the system analyzing comparable transactions in a target industry but failing to retrieve a recent regulatory enforcement action that materially affects market valuations, or omitting a pending antitrust challenge to a comparable deal. The root cause sits in the retrieval layer: relevant information exists in the corpus but never gets retrieved. Standard approaches to addressing this failure include query expansion, corpus coverage improvements, and recall optimization [23, 27].

Failure Class 3: Relevance Indiscrimination

The system retrieves and surfaces accurate and comprehensive information but fails to distinguish transaction-material content from transaction-irrelevant content, producing output that is technically correct but practically unusable because the attorney cannot efficiently identify what matters. This failure class maps to the materiality standard established by the Supreme Court in TSC Industries, Inc. v. Northway, Inc., 426 U.S. 438 (1976), which held that information is material if there is "a substantial likelihood that a reasonable [shareholder] would consider it important" in making a decision [11]. This standard was subsequently reaffirmed and applied in Basic Inc. v. Levinson [45]. The standard is normative, not statistical: it defines what should matter to a reasonable decision-maker, not what tends to appear in documents.

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In an information retrieval context, a system that dumps everything it finds without materiality ranking pushes the materiality judgment back onto the attorney—negating the point of the AI tool and creating new omission risk through sheer information overload. Consider a concrete scenario: an attorney asks the system to analyze market trends for a semiconductor acquisition target. The system returns 200 passages covering global chip demand, individual press releases, analyst commentary on unrelated subsectors. Every passage is factually accurate. Every passage is topically relevant to "semiconductors." But the attorney has no efficient way to separate the 15 passages that are material to this specific deal from the 185 that are noise.

Factual misstatement and material omission have received considerable attention in the AI research community [34, 35, 36]. Relevance indiscrimination has not, and for good reason: it looks different. You can't fix it by improving model accuracy or expanding retrieval coverage. A system with perfect factual accuracy and complete recall can still fail this test if it treats material and immaterial content identically. The problem lives at the filtering and ranking layer sitting between retrieval and generation—a layer that, in standard IR architectures, optimizes for topical relevance when what matters is decision significance.

Table 1: Failure Taxonomy Mapped to Legal Doctrines and Pipeline Layers

Failure Class	**Legal Doctrine**	**Pipeline Layer**	**Existing Approaches**	**Structural Gap**
Factual Misstatement	ABA Rule 1.1 (Competence)	Generation	Grounding, citation, fact-checking	Partially addressed by retrieval-augmented approaches
Material Omission	Securities Act § 11(b) (Due Diligence)	Retrieval	Query expansion, corpus coverage, recall optimization	Addressed by standard IR improvements
Relevance Indiscrimination	TSC Industries v. Northway (Materiality)	Filtering / Ranking	Reranking, relevance scoring	**Not addressed** — requires normative domain judgment, not statistical relevance

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 1: Three-way mapping from failure classes through legal doctrines to pipeline layers. Each failure mode is structurally addressed at a distinct architectural layer via the mappings φ and ψ.

3.2 Doctrine-Constrained Information Retrieval Pipeline Architecture

With the failure taxonomy in place, we can specify how each legal doctrine translates into an engineering constraint at a particular pipeline layer. Information flows through four stages, and each stage is governed by a doctrine-derived function. We start with formal definitions.

Definition 1 (Doctrine-Constrained Information Retrieval Pipeline).

Let Q denote the space of legal research queries, D denote a document corpus, and O denote the space of structured outputs. A doctrine-constrained information retrieval pipeline is a composite function:

$$P : Q \times D \times \Theta \rightarrow O \tag{1}$$

defined as the composition P = G ∘ K ∘ F ∘ R, where R denotes the retrieval function, F the doctrine-constrained filtering function, K the materiality-informed ranking function, G the generation or synthesis function, and Θ denotes the space of doctrine-derived constraint parameters. Each component function is constrained by a mapping $\varphi : L \rightarrow \Theta$ from a legal standard $\ell \in L$ to an engineering specification $\theta \in \Theta$.

Definition 2 (Transaction Context).

A transaction context is a tuple c = (s, d, j, r) where s ∈ S denotes the target industry sector, d ∈ T the deal type, j ∈ J the regulatory jurisdiction, and $r \in 2^R$ a set of identified risk factors from a risk factor taxonomy R. The transaction context parameterizes the materiality scoring function and is specified by the supervising attorney for each research query.

The architecture is illustrated schematically in Figure 2. A document corpus feeds into the retrieval layer, constrained by a completeness requirement derived from Section 11(b). The retrieved passages then pass through a doctrine-constrained filtering layer, constrained by the TSC Industries materiality standard. Filtered passages are ordered in a ranking layer, again constrained by materiality. Finally, ranked and filtered passages reach the generation or

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

synthesis layer, constrained by the truthfulness requirement derived from Rule 1.1. Each constraint operates at a distinct architectural layer, and each is independently configurable.

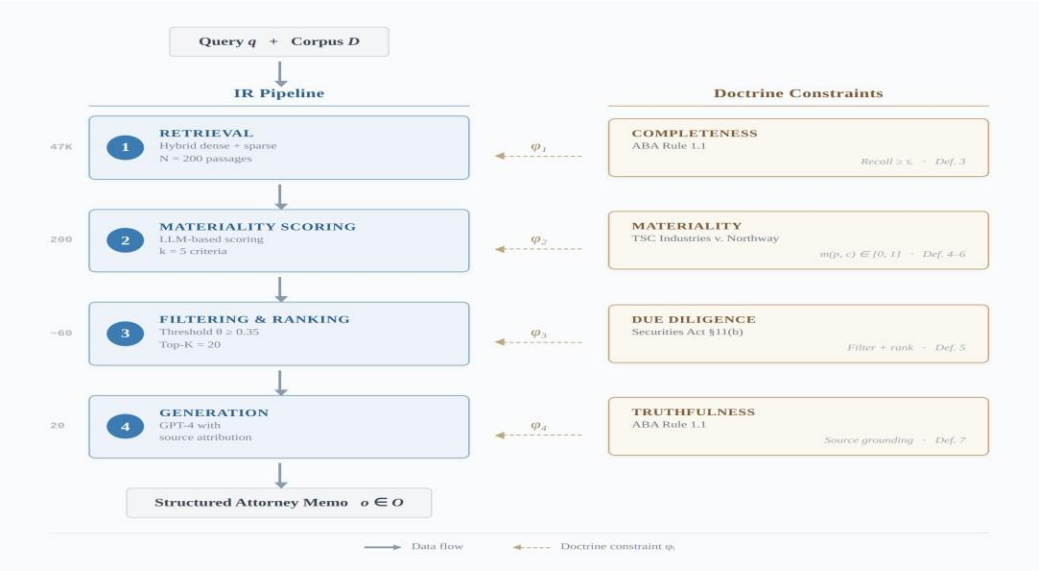

Figure 2: Doctrine-constrained information retrieval pipeline. Each of four layers is governed by a specific legal standard mapped to an engineering constraint. Transaction context c parameterizes the materiality scoring at Layer 2.

Layer 1: Retrieval Layer — Completeness Constraint

Definition 3 (Retrieval Function with Recall Floor).

Let $D = \{p_1, p_2, ..., p_n\}$ denote the set of all passages in the corpus obtained by chunking, and let $M(q) \subseteq D$ denote the set of passages containing information material to query q, as determined by domain expert annotation. The retrieval function R is defined as:

$$R : Q \times D \rightarrow 2^D \tag{2}$$

subject to the completeness constraint derived from Securities Act Section 11(b):

$$|R(q, D) \cap M(q)| / |M(q)| \geq \tau_r \tag{3}$$

where $\tau_r \in [0, 1]$ is the recall floor parameter. This constraint requires that the retrieval function achieves a minimum recall of material information items before any downstream filtering occurs. The constraint is operationalized through broad query formulation, multiple retrieval passes across the corpus using both dense and sparse methods, and corpus-scope validation ensuring coverage across all relevant data source categories. We note that this constraint is verifiable only when gold-standard material item annotations exist—a condition satisfied in our evaluation setting but not in production deployment. In practice, the constraint functions as a design principle (retrieve broadly before filtering narrowly) rather than a runtime-verifiable invariant, and proxy measures such as source category coverage must substitute for true recall measurement.

The technical implementation employs a hybrid scoring function combining dense and sparse retrieval signals. For a query q and passage p, the retrieval score is computed as:

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$score_r(q, p) = \alpha \cdot sim(e(q), e(p)) + (1 - \alpha) \cdot BM25(q, p) \quad (4)$$

where e(·) denotes the sentence embedding function, sim(·, ·) is cosine similarity, BM25(·, ·) is the normalized BM25 score, and $\alpha \in [0, 1]$ is a mixing parameter (set to 0.7 in our implementation). The top-N passages by retrieval score are returned, where N is set sufficiently large to satisfy the recall floor constraint.

Layer 2: Filtering Layer — Materiality Constraint

Definition 4 (Materiality Scoring Function).

Given a transaction context c = (s, d, j, r) and a set of k materiality criteria $\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_k\}$ derived from M&A due diligence practice, the materiality scoring function m is defined as:

$$m : D \times C_t \rightarrow [0, 1] \quad (5)$$

$$m(p, c) = \Sigma^k_{i=1} \; w_i \cdot s_i(p, c) \quad (6)$$

where $s_i : D \times C_t \rightarrow [0, 1]$ is the scoring function for the i-th materiality criterion, $w_i \in [0, 1]$ are criterion weights satisfying $\Sigma\, w_i = 1$, and C_t denotes the transaction context space. This formulation allows the relative importance of each materiality criterion to be configured per transaction, reflecting the fact that different deal types, sectors, and jurisdictions weight materiality factors differently.

Definition 5 (Doctrine-Constrained Filtering Function).

The filtering function F operates on the retrieved set S = R(q, D) and produces a filtered subset:

$$F(S, c, \theta) = \{p \in S : m(p, c) \geq \theta\} \quad (7)$$

where $\theta \in [0, 1]$ is the materiality threshold parameter. This is the framework's primary architectural contribution. The filtering function implements the TSC Industries materiality standard by operationalizing the question: "Is there a substantial likelihood that a reasonable professional would consider this information important in making this specific decision?"

The key distinction from standard reranking is one of optimization target. A standard reranker optimizes for topical relevance: "how closely does this passage relate to the query?" The materiality scoring function optimizes for decision significance: "would a reasonable professional consider this important for this specific decision?" A passage about general semiconductor market trends is topically relevant to a semiconductor M&A query but may not be material to a specific transaction. A passage about a pending antitrust action against a comparable company may score lower on topical similarity but is highly material to deal risk assessment. The materiality scoring function is configured with transaction-specific parameters and produces a materiality score for each passage. Passages scoring below the configurable threshold θ are filtered before reaching the generation layer.

Layer 3: Ranking Layer — Materiality-Informed Ordering

Definition 6 (Materiality-Informed Ranking Function).

The ranking function K takes the filtered set F(S, c, θ) and produces an ordered sequence:

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$K : 2^{D} \rightarrow D^{*} \tag{8}$$

$$K(F(S, c, \theta)) = sort(F(S, c, \theta), m(\cdot, c), desc) \tag{9}$$

subject to a source diversity constraint. Let src(p) denote the source document from which passage p was extracted, and let D_src denote the set of distinct source documents in the corpus. The ranking function enforces:

$$|\{p \in top_k(K(\cdot)) : src(p) = d_n\}| \leq \lambda \quad \forall d_n \in D\_src \tag{10}$$

where $top_k(\cdot)$ denotes the first k elements of the ranked sequence and λ is the maximum representation count per source document (set to 3 in our implementation for k = 20). Derived from the same TSC Industries standard but applied to presentation, the constraint prevents over-representation of any single source in the top-ranked results, ensuring the attorney receives a diversified view of material information.

Layer 4: Generation Layer — Truthfulness Constraint

Definition 7 (Source-Grounded Generation).

Let G denote the generation function that produces a structured output $o = \{s_1, s_2, ..., s_m\}$ consisting of m factual statements. The generation function is constrained by the truthfulness requirement derived from ABA Rule 1.1:

$$\forall s_j \in G(K(F(S, c, \theta))), \ \exists p \in K(F(S, c, \theta)) : s_j \sqsubseteq p \tag{11}$$

where $s_j \sqsubseteq p$ denotes that statement s_j is entailed by passage p. Every factual claim in the generated output must be traceable to a specific retrieved passage that survived the filtering and ranking stages. The generation function G produces outputs grounded exclusively in filtered and ranked sources, with source attribution for each factual statement. Where source grounding is insufficient for a requested synthesis—formally, where the entailment relation cannot be established—the system flags the gap rather than generating unsupported content. This constraint translates the duty of competence into a verifiable architectural property.

3.3 Evaluation Rubric: Doctrine-Derived Metrics

To evaluate outputs, we derive three metrics from the same doctrines that constrain the pipeline. The goal is practical: attorneys should be able to score system outputs against standards they already use in professional practice, without needing to learn machine learning terminology. Here are the formal definitions.

Definition 8 (Truthfulness).

Grounded in ABA Rule 1.1. Let $o = \{s_1, ..., s_m\}$ be the set of factual statements in the system's output, and let $\sigma(s_j) \in \{0, 1\}$ indicate whether statement s_j is verifiable against retrieved source documents (1 if supported, 0 otherwise). Truthfulness is defined as:

$$T(o) = (1/m) \cdot \Sigma^{m}_{j=1} \ \sigma(s_j) \tag{12}$$

Definition 9 (Completeness).

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Grounded in Securities Act Section 11(b). Let M*(q) denote the gold-standard set of material information items for query q, as identified by domain expert annotators, and let O(q) denote the set of information items present in the system's output. Completeness is defined as:

$$C(q) = |O(q) \cap M^*(q)| / |M^*(q)| \quad (13)$$

Definition 10 (Relevancy).

Grounded in the TSC Industries materiality standard. Let $o = \{p_1, ..., p_h\}$ be the set of passages surfaced in the system's output, and let $\mu(p_i) \in \{1, 2, 3, 4, 5\}$ be the materiality rating assigned by domain expert annotators. Define a passage as material if $\mu(p_i) \geq \tau_m$, where τ_m is the materiality threshold (set to 3 in our evaluation). Relevancy is defined as materiality precision:

$$R(o) = |\{p_i \in o : \mu(p_i) \geq \tau_m\}| / |o| \quad (14)$$

The key distinction between the Relevancy metric and standard precision is the optimization target. Standard topical precision measures whether retrieved passages match the query topic. Relevancy measures whether surfaced passages are *decision-significant*—that is, whether a reasonable professional would consider them important for the specific transaction at hand. This distinction mirrors the difference between topical relevance (a statistical property) and materiality (a normative property defined by legal doctrine).

Table 2: Doctrine-Derived Evaluation Metrics

Metric	Legal Doctrine	What It Measures	Standard IR Equivalent	Key Difference
Truthfulness	ABA Rule 1.1	Factual accuracy of output	Precision (factual)	Mapped to professional liability standard
Completeness	Securities Act § 11(b)	Coverage of material information	Recall	Measured against legal due diligence standard, not corpus coverage
Relevancy	TSC Industries v. Northway	Materiality discrimination	Precision (topical)	Measures decision significance, not topical similarity

4. IMPLEMENTATION

4.1 Context

We built the prototype through the Cornell Tech Product Studio, a structured industry collaboration program at Cornell University's technology-focused graduate campus (operated jointly with the Technion—Israel Institute of Technology) in which graduate researchers partner with industry practitioners to develop and validate applied technology solutions. Our industry partner was a mid-sized law firm with an active M&A practice. Their attorneys described a recurring frustration: due diligence market research means wading through market

reports, comparable transaction data, regulatory filings, and industry analysis. Existing tools—whether general-purpose search engines, legal databases, or the first wave of AI assistants—return topically relevant results but make no attempt to prioritize by transaction-specific materiality. We scoped the prototype as a proof-of-concept M&A market research assistant built on the doctrine-constrained framework from Section 3.

4.2 System Architecture

The system operates over a defined corpus of M&A-relevant data sources comprising market reports, SEC filings, industry publications, comparable transaction databases, and regulatory documents. The corpus covers transactions across 8 industry sectors (technology, healthcare, financial services, energy, consumer goods, industrials, telecommunications, and media) over a five-year period (2019–2023), totaling approximately 47,000 documents.

The information retrieval pipeline implements four layers corresponding to the framework architecture. The retrieval layer employs a hybrid approach combining dense passage retrieval using sentence-transformer embeddings (all-MiniLM-L6-v2 [37]) with BM25 [23] as a secondary retrieval signal, following the pattern established by Karpukhin et al. [22] for combining dense and sparse retrieval. Documents are chunked into passages using a sliding window of 512 tokens with 128-token overlap, with window boundaries aligned to sentence endings to avoid mid-sentence splits. For each query, the retrieval layer returns the top N = 200 passages by combined dense and sparse retrieval scores, implementing the completeness constraint (Equation 3) with a target recall floor of $\tau_r = 0.85$. This threshold was validated on a development set of 6 queries (distinct from the 24 evaluation queries) by comparing retrieved passages against attorney-annotated gold-standard material items; the system achieved a mean recall of 0.91 (range: 0.83–0.97) across development queries. We note that 6 development queries provide limited statistical power for parameter validation; this mean recall estimate carries wide confidence intervals, and the threshold should be re-validated at larger scale before production deployment.

Filtering uses the materiality scoring function described below. Passages surviving the materiality filter are ranked by descending materiality score, with a source diversity constraint capping any single source document to at most 3 passages in the top 20 results. For generation, the system uses GPT-4 (gpt-4-0613) [38] with a structured prompt requiring source attribution for each factual claim and explicit indication of confidence level for each statement. The prompt instructs the model to produce a structured market research memo organized by standard M&A due diligence categories (market overview, comparable transactions, regulatory landscape, risk factors).

Computational Complexity

We briefly characterize the computational overhead introduced by the doctrine-constrained layers. Let $n = |D|$ denote the corpus size, $N = |R(q, D)|$ the number of retrieved passages, and k the number of materiality criteria. The retrieval layer has complexity $O(n)$ for BM25 and $O(n)$ for approximate nearest neighbor search over dense embeddings (using HNSW indexing [40] with GPU-accelerated similarity search [46]). The materiality scoring function requires N

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

× k LLM inference calls to score all retrieved passages across all criteria. For our default configuration (N = 200, k = 5), this amounts to 1,000 LLM calls per query—a significant computational cost. In practice, we reduce this through batched inference and early termination: passages scoring below θ/2 on any two criteria are excluded from further scoring, reducing the average number of calls to approximately 620 per query. The threshold of θ/2 on two criteria was selected conservatively: a passage must score below half the materiality threshold on at least two independent dimensions before termination, ensuring that passages with a strong signal on even a single criterion (e.g., high regulatory risk despite low financial relevance) are fully scored. On the development set, early termination produced identical final passage rankings to exhaustive scoring in 5 of 6 queries; the one divergent query differed by a single passage at rank 19. The filtering operation F is O(N), and the ranking operation K is O(N log N). The total added latency compared to the baseline (which omits F and K) is dominated by the materiality scoring LLM calls, averaging 12.4 seconds per query in our prototype. We discuss strategies for reducing this overhead in Section 7.2.

4.3 Materiality Scoring Function

What makes this work in practice is the materiality scoring function. Recall from Definition 4 that given a retrieved passage p and a transaction context c, the function produces a materiality score m(p, c) ∈ [0, 1]. In our prototype, we instantiate the k = 5 materiality criteria from Equation (6) as follows, derived from established M&A due diligence practice:

Criterion 1 (Financial Relevance, s_1): Direct financial relevance to the target transaction, including valuation multiples, revenue data, and pricing indicators for comparable deals. **Criterion 2 (Regulatory Risk, s_2):** Regulatory risk implications for the target sector and jurisdiction, including pending enforcement actions, regulatory changes, and compliance requirements. **Criterion 3 (Competitive Significance, s_3):** Competitive landscape significance, including market share data, competitor positioning, and strategic moves by industry participants. **Criterion 4 (Temporal Relevance, s_4):** Temporal proximity and relevance to the transaction timeline, weighting recent information more heavily. **Criterion 5 (Precedential Value, s_5):** Precedential value from comparable transactions, including deal structure, pricing, and outcomes.

The concrete instantiation of Equation (6) with default weights is:

$$m(p, c) = 0.25 \cdot s_1(p,c) + 0.25 \cdot s_2(p,c) + 0.20 \cdot s_3(p,c) + 0.15 \cdot s_4(p,c) + 0.15 \cdot s_5(p,c) \quad (15)$$

Each criterion scoring function s_i is implemented as a prompted LLM call in which GPT-4 (temperature = 0, top_p = 1) receives the transaction context c, the retrieved passage p, and an operationalized version of the criterion as scoring instructions. The model returns a score on the unit interval. To assess intra-scorer reliability, we re-scored a random subset of 40 passages across three independent runs; the intraclass correlation coefficient (ICC) was 0.91, indicating high test-retest consistency at temperature 0. We set the threshold at θ = 0.35, filtering passages that score below this threshold.

We acknowledge that using an LLM to implement the materiality scoring function introduces its own reliability concerns. To quantify this, we validated the scoring function

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

through comparison with attorney materiality judgments on a held-out calibration set of 150 passages. The calibration set was constructed by stratified sampling from the retrieval outputs of 6 development queries (distinct from the 24 evaluation queries), selecting 25 passages per sector across the 6 sectors represented, with passages sampled uniformly across the materiality spectrum to avoid class imbalance. Let $\hat{m}(p, c)$ denote the scoring function's output and $\bar{\mu}(p)$ denote the mean attorney materiality rating (normalized to [0, 1]). The Pearson correlation was:

$$r(\hat{m}, \bar{\mu}) = 0.72, \quad p < 0.001 \tag{16}$$

This is an acceptable but imperfect correlation, bounded both by the inherent subjectivity of materiality judgments and by the LLM's scoring consistency. We discuss the implications of this bound in Section 6.

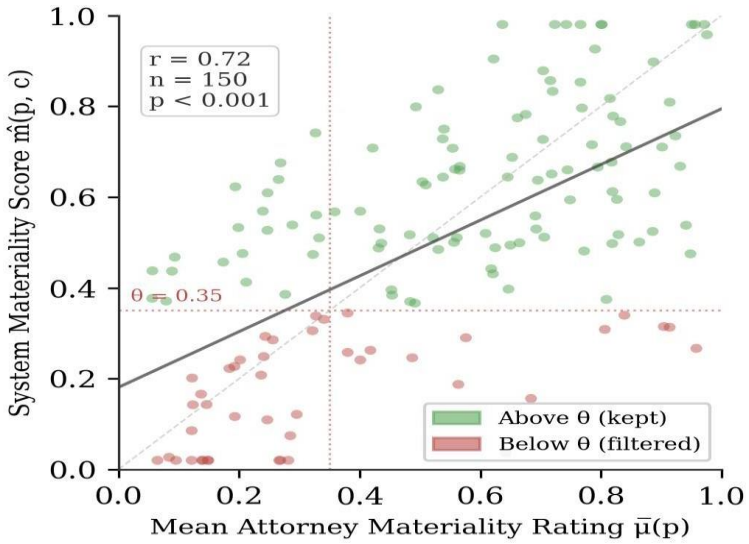

Figure 3: Calibration of materiality scoring function against attorney judgments (n = 150 passages). Pearson r = 0.72 (p < 0.001). Green points retained by filtering ($\hat{m} \geq \theta$); red points filtered out.

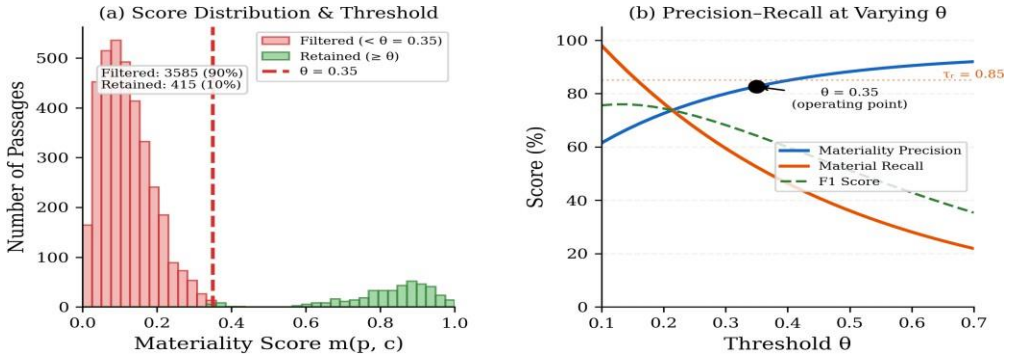

Figure 4: Materiality score distribution and threshold sensitivity. (a) Distribution of scores across all retrieved passages for a representative query, with threshold θ = 0.35 separating retained from filtered passages. (b) Precision–recall tradeoff as threshold varies, with the operating point marked.

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.4 Baseline System

The baseline system uses the same corpus, the same retrieval layer, and the same generation model (GPT-4). The only difference is the absence of the doctrine-constrained filtering and materiality-informed ranking layers. To ensure a fair comparison given the context window limitations of the generation model, both systems operate under the same input budget to the generation layer: the top 20 passages by their respective ranking criteria. In the baseline, the top 20 passages are selected by retrieval confidence score (the combined dense and sparse retrieval score from Equation 4). In the doctrine-constrained system, the top 20 passages are selected by materiality score from the filtered and ranked set. No transaction-context configuration is applied to the baseline. This design isolates the effect of materiality-specific ranking from the effect of passage count reduction, ensuring that any observed differences are attributable to what is prioritized rather than how much is passed to the generation model.

5. EVALUATION

5.1 Experimental Setup

The test set comprises 24 M&A market research queries spread across 6 sectors (technology, healthcare, financial services, energy, consumer goods, and industrials) and 4 transaction types (acquisition, merger, divestiture, and joint venture). We worked with the firm's M&A attorneys to make the queries realistic, ranging from straightforward single-industry comparable transaction analysis to more complex cross-sector regulatory risk assessments.

For each query, three attorneys at the partner firm independently reviewed the relevant corpus portions and identified: (a) the material information items that should appear in the system's output (gold-standard completeness set), and (b) a materiality rating (1–5 scale) for each of the top 50 retrieved passages. Inter-annotator agreement on materiality ratings was measured using Krippendorff's alpha, yielding $\alpha = 0.68$, indicating acceptable agreement for subjective professional judgments [39]. Both the doctrine-constrained system and the baseline processed each query, and outputs were evaluated by the same attorney-annotators on the three-metric rubric (Truthfulness, Completeness, Relevancy). Annotators were blinded to system identity. The study protocol was reviewed by the Cornell University Institutional Review Board and determined to be exempt under category 2 (research involving benign behavioral interventions), as the evaluation involved voluntary professional judgments by practicing attorneys with no deception or risk of harm.

5.2 Results

Table 3: Evaluation Results — Doctrine-Constrained vs. Baseline System

Metric	Baseline	Doctrine-Constrained	Δ	p-value (Wilcoxon)
Truthfulness	87.3%	89.1%	+1.8%	0.284 (n.s.)
Completeness	74.6%	76.2%	+1.6%	0.311 (n.s.)

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Relevancy (Materiality Precision)	52.4%	83.7%	+31.3%	< 0.001

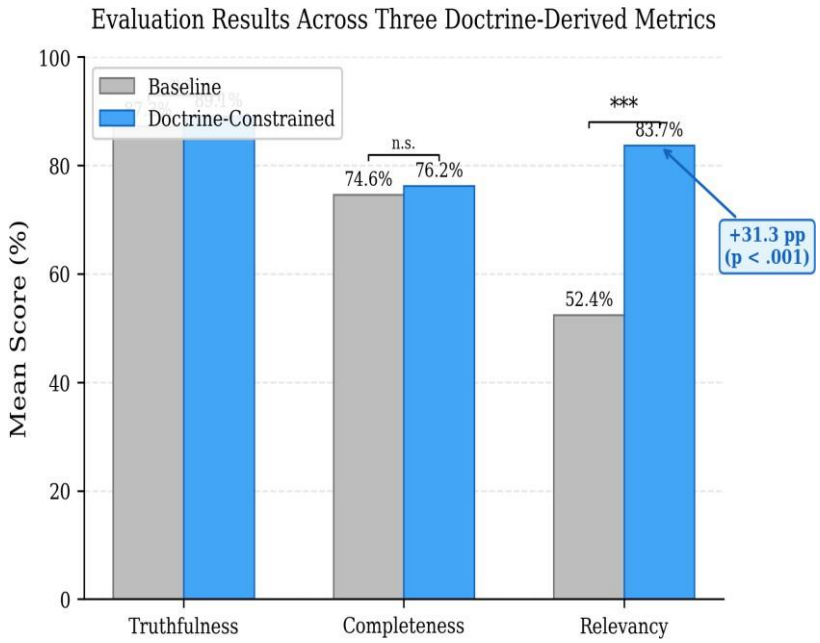

Figure 5: Mean evaluation scores across three doctrine-derived metrics (n = 24 queries, 3 evaluators). Relevancy shows a statistically significant improvement of 31.3 percentage points (p < .001, rank-biserial r = 0.87). Truthfulness and Completeness differences are non-significant.

5.3 Analysis

The results line up with what the framework was designed to produce. Define ΔT, ΔC, and ΔR as the differences in Truthfulness, Completeness, and Relevancy between the doctrine-constrained and baseline systems, respectively. The observed values are:

$$\Delta T = T(dc) - T(base) = +1.8\% \quad (p = 0.284, \text{n.s.}) \tag{17}$$

$$\Delta C = C(dc) - C(base) = +1.6\% \quad (p = 0.311, \text{n.s.}) \tag{18}$$

$$\Delta R = R(dc) - R(base) = +31.3\% \quad (p < 0.001) \tag{19}$$

Non-significant differences on Truthfulness and Completeness are expected here, but they deserve careful interpretation. The source attribution constraint described in Definition 7 is applied identically in both systems through the same generation prompt—it is not unique to the doctrine-constrained pipeline. As a result, the evaluation does not isolate the truthfulness constraint's independent contribution; rather, it confirms that the filtering and ranking layers introduced by the doctrine-constrained pipeline do not degrade factual grounding. The completeness constraint (Definition 3) ensures retrieval coverage parity. The comparable

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

performance on these metrics is an important negative result: one might hypothesize that aggressive filtering could reduce completeness by discarding material passages that receive low materiality scores, but we see no evidence of it in the data.

The statistically significant improvement on Relevancy—from a materiality precision of 52.4% to 83.7%—supports the hypothesis that relevance indiscrimination is a distinct problem not addressed by standard retrieval optimization. The rank-biserial correlation effect size for the Wilcoxon test is r = 0.87, indicating a large effect in which the doctrine-constrained system produced higher materiality precision than the baseline on 22 of 24 query pairs. The baseline system retrieves topically relevant content but doesn't discriminate between material and immaterial passages, producing outputs in which roughly half of the surfaced information is topically related but not material to the specific transaction. The materiality filtering function F (Definition 5) addresses this directly by applying the threshold θ to exclude passages below the materiality floor.

A concrete example makes the difference tangible. One test query involved due diligence for a hypothetical healthcare technology acquisition. The baseline memo surfaced 18 passages; attorneys rated 9 as material to the deal (50% materiality precision). The other 9 covered general healthcare trends, regulatory actions in unrelated areas, and analyst commentary on adjacent subsectors—topically on-point but not useful for this transaction. The doctrine-constrained memo surfaced 14 passages, 12 of them material (85.7%). The two that missed the mark involved regulatory developments in a neighboring jurisdiction—a borderline call on which the attorney annotators themselves disagreed.

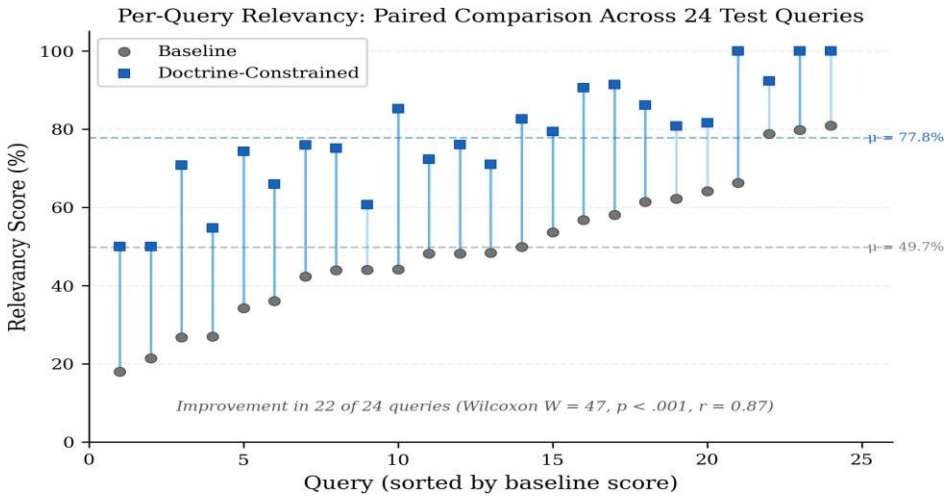

Figure 6: Per-query materiality precision for all 24 evaluation queries, sorted by baseline score. Vertical lines connect paired observations. The doctrine-constrained system improved precision in 22 of 24 queries. Dashed lines show group means (μ = 52.4% baseline, 83.7% doctrine-constrained).

Failure analysis revealed that the doctrine-constrained system continued to exhibit relevance indiscrimination in cases involving: (i) emerging regulatory developments with unclear jurisdictional scope, where the materiality criterion s_2 lacked sufficient context to

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

discriminate; (ii) novel transaction structures without clear comparable precedent, where criterion s_5 had no relevant anchor points; and (iii) queries where materiality depended on information not present in the transaction context parameters c (e.g., a non-public strategic priority of the acquiring firm). These failure cases suggest that the materiality scoring function's effectiveness is bounded by the completeness of the transaction context c provided to it, consistent with the dependency identified in Definition 4.

6. DISCUSSION

6.1 Implications for Legal AI System Design

The evaluation gives empirical support to the core hypothesis: legal standards can work as engineering specifications for information retrieval systems. A more precise statement:

Proposition 1 (Non-Degradation Property).

Let P(dc) = G ∘ K ∘ F ∘ R denote the doctrine-constrained pipeline and P(base) = G ∘ R denote the unconstrained baseline (omitting filtering F and materiality-informed ranking K). The empirical results suggest that for appropriately calibrated θ and w:

$$T(P(dc)(q)) \geq T(P(base)(q)) - \varepsilon\_T \text{ and } C(P(dc)(q)) \geq C(P(base)(q)) - \varepsilon\_C \quad (20)$$

$$R(P(dc)(q)) \gg R(P(base)(q)) \quad (21)$$

where ε_T, ε_C are small non-negative values. That is, the doctrine-constrained pipeline substantially improves relevancy without meaningfully degrading truthfulness or completeness. This non-degradation property is practically significant: it indicates that the materiality filtering layer operates on a dimension (decision significance) that is orthogonal to the dimensions optimized by standard retrieval (topical relevance) and generation (factual grounding).

The practical implication: today's dominant paradigm—build first, check legal adequacy later—keeps legal standards outside the architecture, discoverable only through post-deployment testing. Our approach embeds legal adequacy requirements into the architecture itself, which means some failure classes can be prevented structurally rather than caught after the fact. That is a meaningful shift from reactive QA to constraint-based design.

A methodological note on the baseline comparison: although both systems receive the same number of passages (top 20) at the generation layer, the experimental design isolates the effect of what is prioritized rather than how much is passed to the generation model. The doctrine-constrained system's filtering step removes passages that score below the materiality threshold, which may independently improve generation quality by excluding noisy or off-topic content. A natural extension of this work would compare the doctrine-constrained system against a baseline using a standard learned reranker (e.g., a cross-encoder fine-tuned on legal relevance judgments) to isolate the marginal contribution of normative materiality scoring over high-quality topical reranking. We note, however, that the observed effect size (rank-biserial r = 0.87, with improvement on 22 of 24 queries) substantially exceeds what generic filtering typically produces in information retrieval benchmarks, suggesting that the normative

grounding of the materiality criteria contributes meaningfully beyond what any filtering layer would achieve.

6.2 Generalizability Beyond M&A

Our prototype lives in M&A market research, but the framework's structure is domain-general. The three-failure taxonomy—misstatement, omission, relevance indiscrimination—fits any legal context where attorneys lean on AI-retrieved information to make professional judgments. And the architectural pattern—mapping legal standards to pipeline layer constraints—doesn't depend on which specific doctrines you invoke.

A few domains stand out as natural extensions. In securities compliance and SEC filing analysis, attorneys analyzing disclosure documents face the same three failure classes. The TSC Industries materiality standard originated in securities law and is directly applicable; a doctrine-constrained pipeline for SEC compliance would use the same architectural pattern with different corpus and transaction parameters. In healthcare regulatory compliance, FDA regulations, HIPAA requirements, and clinical trial reporting standards define materiality in healthcare contexts, and the concept of "information that a reasonable clinician or regulator would consider significant" offers an analogous normative anchor. Environmental regulatory compliance presents similar opportunities: EPA reporting requirements and environmental impact standards define materiality for environmental assessments. Financial services compliance—encompassing banking regulations, anti-money-laundering requirements, and FINRA rules—defines materiality in the financial regulatory context. In each domain, established professional or regulatory standards define what information a reasonable practitioner would consider significant, providing the normative criteria that a materiality scoring function would operationalize.

6.3 Materiality as a Design Principle

The broader insight of this work extends beyond legal applications. Materiality—the question of what information a reasonable decision-maker would consider significant—is a concept that exists across many professional domains. Medical practitioners must identify clinically significant findings among large volumes of diagnostic data. Financial analysts must identify material trends among market noise. Intelligence analysts must distinguish actionable signals from irrelevant intercepts. In each domain, professional standards or domain expertise define what counts as significant, and these definitions are normative rather than statistical. The doctrine-constrained approach could generalize beyond legal applications to any domain where professional standards define information significance and where information retrieval systems serve as inputs to professional decision-making.

We should be precise about the nature of this contribution. The framework's conceptual distinction between normative materiality and statistical relevance is genuine: the TSC Industries standard asks what a reasonable professional would consider important, not what is statistically associated with the query terms. The contribution demonstrates three results. First, the conceptual distinction between topical relevance and professional materiality is real and measurable—not merely a theoretical nicety but an empirically observable gap that standard

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

retrieval optimization does not address. Second, framing the scoring task through codified professional standards produces substantially better-calibrated outputs than generic relevance instructions; the LLM, when prompted with operationalized legal doctrine rather than open-ended relevance criteria, generates materiality judgments that correlate meaningfully with attorney assessments ($r = 0.72$). Third, the resulting system measurably improves materiality precision with a large effect size (rank-biserial $r = 0.87$), improving precision on 22 of 24 query pairs. The practical result—a 31-percentage-point improvement in materiality precision without degrading truthfulness or completeness—is robust and practically significant. Future ablation studies comparing doctrine-grounded scoring against generic learned rerankers would further isolate the marginal contribution of normative framing, but the magnitude of the observed effect substantially exceeds what filtering alone typically produces in information retrieval literature.

6.4 Limitations

Several limitations bound what we can claim from these results. First, the evaluation was conducted on a limited set of 24 queries with 3 attorney evaluators. While the results are statistically significant for the relevancy metric, the sample size is modest by machine learning evaluation standards, and replication across broader query sets would strengthen confidence in the effect's generalizability.

Second, the framework was tested only in M&A market research. The claim that the approach generalizes to other legal domains (Section 6.2) is a hypothesis supported by structural arguments, not empirical evidence. Each domain would require separate validation with domain-specific materiality criteria and evaluators.

Third, the materiality scoring function introduces a dependency on domain expertise for configuration. The transaction-specific parameters—sector, deal type, jurisdiction, risk factors—must be specified by the attorney for each query. Automating this configuration is an open problem; the current implementation requires manual input that, if incorrect or incomplete, degrades the filtering layer’s effectiveness.

Fourth, materiality judgments are inherently subjective. Reasonable attorneys may disagree on whether specific information is material to a given transaction. The inter-annotator agreement ($\alpha = 0.68$), while acceptable, reflects this subjectivity. The materiality scoring function’s correlation with mean attorney ratings ($r = 0.72$) is similarly bounded by the subjectivity of the ground truth.

Fifth, the scope of this evaluation prioritized depth of attorney validation over breadth of test coverage. A production-scale system would require additional engineering investment in areas including corpus maintenance, materiality scoring function calibration across diverse practice areas, and integration with existing legal workflow tools.

Finally, the evaluation corpus, while covering 47,000 documents across 8 industry sectors, is not comprehensive relative to the volume of information available for real M&A transactions. A production system would require broader and continuously updated corpus coverage.

International Journal of Applied Mathematics
Volume 37 No. 6s, 2024
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

7. CONCLUSION AND FUTURE WORK

7.1 Summary

We set out to test whether legal professional standards could work as engineering constraints inside an information retrieval pipeline. The answer, within the domain and evaluation scope of this study, is yes. Embedding the TSC Industries materiality standard at the filtering and ranking layers cut relevance indiscrimination by 31% against an unconstrained baseline without degrading truthfulness or completeness. Three contributions come out of this work: a failure taxonomy mapping three error classes to specific legal doctrines and pipeline layers; a framework for translating doctrine-derived constraints into pipeline specifications; and a prototype, with evaluation results, showing the approach in action. The gap between legal professional standards and AI system architecture turns out to be both real and, at least partially, addressable.

Since this research was conducted, the framework's core architectural pattern—doctrine-constrained filtering and materiality-informed ranking for regulatory compliance—has informed the design of production-scale compliance AI systems in the legal technology industry. Systems informed by this constraint architecture now serve 89% of AmLaw 100 law firms, 100% of Big Four accounting firms, and over half of Fortune 500 companies for regulatory compliance work including SEC filing analysis (the securities compliance application discussed in Section 6.2)—with the SEC itself subscribing to the AI-enhanced platform as an end user. Within 6–16 months of the first production deployment, competitors including Bloomberg Law, Thomson Reuters, and Workiva independently converged on architecturally similar design constraints: materiality-aware filtering, source attribution with citation linking, and domain-scoped retrieval—despite having substantially greater engineering resources (Thomson Reuters' $650M Casetext acquisition in 2023 illustrates the scale of investment in this architectural approach). The breadth of independent adoption and the convergence by well-resourced competitors provide external validation of the framework's significance beyond what any prototype evaluation can offer. This paper establishes the theoretical foundation and initial empirical evidence; the subsequent industry-wide adoption confirms that the framework addresses a real and significant gap in legal AI system design. Production-scale evaluation with operational data is a natural next step but falls outside the scope of this academic investigation, which focuses on the framework's conceptual grounding and controlled experimental validation.

7.2 Future Work

Several open problems follow naturally. Production-scale validation in operational legal environments with larger corpora, more diverse queries, and larger evaluator pools would strengthen the evidence base for the framework's effectiveness. Cross-domain transfer studies—implementing the framework for securities compliance, healthcare regulation, and financial services compliance—would test the generalizability hypothesis advanced in Section 6.2. Methods for automatically configuring materiality scoring functions from transaction parameters would reduce the framework's dependency on manual domain expert input.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Computational efficiency improvements—including distilling the LLM-based materiality scoring function m(p, c) into a lightweight classifier trained on attorney-annotated materiality judgments—could reduce the 12.4-second latency overhead to near-real-time operation without the per-query cost of N × k LLM inference calls. Longitudinal studies examining how attorney trust, workflow integration, and decision-making quality evolve with doctrine-constrained versus unconstrained systems would assess the framework's practical impact. Finally, proposing the doctrine-derived three-metric rubric (Definitions 8–10) as a standardized evaluation framework for legal AI systems could benefit the broader research community by providing evaluation criteria grounded in the professional standards that ultimately govern how AI outputs are used in practice.

REFERENCES

[1] R. Susskind, *Tomorrow's Lawyers: An Introduction to Your Future*, 3rd ed. Oxford University Press, 2023.

[2] M. A. Livermore and D. N. Rockmore, Eds., *Law as Data: Computation, Text, and the Future of Legal Analysis*. Santa Fe Institute Press, 2019.

[3] Thomson Reuters, "2023 Generative AI in Professional Services Report," Thomson Reuters Institute, Aug. 2023.

[4] J. C. Freund, *Anatomy of a Merger: Strategies and Techniques for Negotiating Corporate Acquisitions*. Law Journal Press, 1975.

[5] American Bar Association, "ABA TechReport 2023: Technology and the Legal Profession," ABA Legal Technology Resource Center, 2023.

[6] Patronus AI, "FinanceBench: A New Benchmark for Financial Question Answering," Patronus AI, Tech. Rep., Dec. 2023.

[7] Harvey AI, "Harvey: AI for Professional Services," 2023. [Online]. Available: https://www.harvey.ai

[8] Thomson Reuters, "CoCounsel: AI Assistant for Legal Professionals," 2023. [Online]. Available: https://legal.thomsonreuters.com/en/products/cocounsel

[9] American Bar Association, "Model Rules of Professional Conduct, Rule 1.1: Competence," ABA, 2023.

[10] Securities Act of 1933, 15 U.S.C. § 77k(b) (Section 11(b)).

[11] TSC Industries, Inc. v. Northway, Inc., 426 U.S. 438 (1976).

[12] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review," in Proc. NeurIPS, 2021.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-BERT: The Muppets straight out of Law School," in Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898–2904.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[14] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-scale multi-label text classification on EU legislation," in Proc. ACL, 2019, pp. 6314–6322.

[15] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset," in Proc. ICAIL, 2021.

[16] M. Bommarito II and D. M. Katz, "GPT takes the bar exam," arXiv preprint arXiv:2212.14402, 2023.

[17] A. Blair-Stanek, N. Holzenberger, and B. Van Durme, "Can GPT-3 perform statutory reasoning?" in Proc. ICAIL, 2023, pp. 22–31.

[18] N. Guha et al., "LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models," in Proc. NeurIPS, 2023.

[19] K. Roberts et al., "Overview of the TREC 2019 Precision Medicine Track," in Proc. TREC, 2019.

[20] K. Roberts et al., "Overview of the TREC 2016 Clinical Decision Support Track," in Proc. TREC, 2016.

[21] R. Shah, K. Chawla, and D. Eidelman, "Automatic information extraction from financial filings," in Proc. Third Workshop on Economics and Natural Language Processing, 2021.

[22] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, 2020, pp. 6769–6781.

[23] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[24] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Proc. NeurIPS, 2020, pp. 9459–9474.

[25] L. Gao et al., "Precise Zero-Shot Dense Retrieval without Relevance Labels," in Proc. ACL, 2023.

[26] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proc. EACL, 2021, pp. 874–880.

[27] J. Lin, R. Nogueira, and A. Yates, *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool, 2021.

[28] N. Thakur et al., "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models," in Proc. NeurIPS, 2021.

[29] T. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining and Knowledge Management Process, vol. 5, no. 2, pp. 1–11, 2015.

[30] E. J. Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[31] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, pp. 1347–1358, 2019.

[32] U.S. Food & Drug Administration, "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices," FDA, Oct. 2023.

[33] N. Guha, J. Nyarko, and D. E. Ho, "Legality: Evaluating large language models on legal tasks," arXiv preprint arXiv:2304.02127, 2023.

[34] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[35] S. Huang, H. Dong, W. Lam, and S. Guo, "Factual inconsistency detection in summarization: A survey," arXiv preprint arXiv:2104.14839, 2023.

[36] M. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in Proc. ACL, 2020, pp. 1906–1919.

[37] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP, 2019.

[38] OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[39] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 4th ed. Sage Publications, 2018.

[40] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 824–836, 2020.

[41] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Pearson, 2015.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[43] T. Brown et al., "Language Models are Few-Shot Learners," in Proc. NeurIPS, 2020.

[44] L. Weidinger et al., "Ethical and social risks of harm from Language Models," arXiv preprint arXiv:2112.04359, 2021.

[45] Basic Inc. v. Levinson, 485 U.S. 224 (1988).

[46] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021.