What doesn't match matters more

CRANE: Calibrated Retrieval with Adversarial Negative Evidence

Donald Murre

Abstract

Retrieval-Augmented Generation (RAG) systems retrieve documents by positive relevance (how similar a document is to a query) and generate answers from whatever scores highest. This paradigm has no mechanism for representing what a document does *not* answer. We argue that this omission is the root cause of the most dangerous RAG failures: confident wrong answers in high-stakes domains.

We present evidence from legal, medical, and financial retrieval that positive-only scoring fails in three systematic ways: (1) it cannot distinguish topically similar but jurisdictionally inapplicable documents (*scope failure*); (2) it cannot distinguish related but categorically different concepts (*sibling failure*); and (3) it cannot distinguish current from superseded versions of a document (*temporal failure*). Each failure mode produces false positives that are invisible to the generation layer.

We derive three properties that any retrieval system adequate for high-stakes domains must satisfy: (1) typed negative evidence representation, capturing specific ways a document fails to answer a query; (2) calibrated retrieval confidence, producing $P(\text{relevant})$ rather than ranking scores; and (3) confidence-gated generation, selecting response strategies based on the type and severity of uncertainty. We define calibrated retrieval formally and propose a three-type taxonomy of negative evidence: *scope*, *sibling*, and *temporal* negatives.

As proof by construction, we present CRANE (**C**alibrated **R**etrieval with **A**dversarial **N**egative **E**vidence), a functional architecture satisfying all three properties. CRANE's mechanisms (typed negative detection, calibrated merge scoring, and confidence-gated generation) produce measurable technical effects: reduced false-positive retrieval, calibrated relevance probabilities, and differentiated generation conditioned on uncertainty type. Other architectures meeting the same requirements would share the same structural advantages. We state four falsifiable predictions and invite empirical evaluation.

1 Introduction

A tax lawyer asks a retrieval-augmented system: "Does Article 15(2) of the Swiss–Germany Double Taxation Agreement (DTA) apply to cross-border remote workers post-2024?" The system retrieves ten chunks, ranks them by similarity, and generates an answer. The answer is confident, well-sourced, and wrong. Two of the retrieved chunks are from Swiss domestic tax law, not from the bilateral treaty. One is from the pre-2024 version of the treaty, which does not address remote workers. The system scored all three highly because they are textually similar to the query. It has no representation of "this document does not apply."

This is not a rare failure. Magesh et al. (2024) report that retrieval-augmented legal tools hallucinate at rates between 17% and 33%, higher in some configurations than unassisted models. Daivam (2025), in an industry case study, documents false-positive rates exceeding 99% in a banking RAG system at a standard 0.7 similarity threshold: a customer asking to cancel a credit card was matched to investment account closure procedures with 88.7% confidence. The higher the confidence, the greater the danger, because a system that is wrong at high confidence is one that no downstream check can catch.

The root cause is architectural. Current retrieval systems model what a document *is about* but have no mechanism for representing what it is *not about*. A retriever assigns a positive relevance score to each document–query pair, with higher scores indicating more relevance: higher means more relevant, lower means less. There is no negative channel. The system cannot say "this document addresses the right topic but the wrong jurisdiction," or "this version has been superseded," or "this concept is related but categorically different from the one the query asks about." In legal reasoning, the distinction has a name. A court does not merely find the most relevant authority; it distinguishes and rules out inapplicable ones. The reasons for exclusion are as much part of the judgment as the reasons for inclusion. Current RAG systems have no equivalent.

We use "adversarial" in this paper in the legal-proceeding sense, i.e., testing claims against counter-evidence, not in the machine-learning sense of adversarial attacks. The analogy is to cross-examination, not to perturbation.

The asymmetry matters because different domains tolerate different error profiles. In web search, a wrong result is an inconvenience. In legal, medical, and financial retrieval, a wrong result that looks right is a liability. Under the current paradigm, RAG systems operate on what a lawyer would recognize as a preponderance-of-evidence standard: the highest-scoring document wins. High-stakes domains need something closer to beyond-reasonable-doubt: rule out the alternatives before asserting the answer.

We make four contributions:

1. **The negative evidence thesis.** We argue, with evidence from legal, medical, and financial domains, that in order to correctly identify what documents cover, retrieval systems for high-stakes applications must model what documents do *not* answer, not only what they do.

2. **A taxonomy of negative evidence types.** We propose a three-type classification of negative retrieval evidence (*scope negatives*, *sibling negatives*, and *temporal negatives*), each

corresponding to a distinct failure mode of positive-only retrieval. Temporal negatives operate at claim-level granularity: individual provisions within a document may have different temporal statuses, a distinction absent from existing version-aware retrieval systems.

3. **The calibrated retrieval requirement.** We define what it means for a retrieval system to be *calibrated* (producing $P(\text{relevant})$ rather than rank-order scores) and derive this requirement from first principles using the Probability Ranking Principle.

4. **Crane as proof by construction.** We present Crane (Calibrated Retrieval with Adversarial Negative Evidence), an architecture that jointly satisfies all three requirements. The properties are general: other architectures satisfying them would share the same advantages.

The running example recurs throughout (see Figure 2 for its embedding-space representation). We begin by tracing how positive-only retrieval became the default (Figure 1), then present evidence that it fails in systematic, predictable ways. From that evidence we derive the properties that any adequate system must satisfy and present Crane as one architecture meeting all three. We address the strongest objections, including the claim that better embeddings or larger context windows make negative evidence unnecessary, and close with falsifiable predictions and a research roadmap.

2 Background: the positive-only paradigm

2.1 From BM25 to dense retrieval

For three decades, the dominant model of text retrieval has rested on one assumption, namely that relevance is a single positive score. The scoring function has changed, from term frequency (BM25) to learned dense representations (Karpukhin et al., 2020), but the assumption has not.

BM25 computes relevance as weighted term overlap. A document scores high if it contains the query's terms, weighted by inverse document frequency. Dense retrievers replace term overlap with embedding proximity: the query and document are encoded as vectors, and relevance is measured by cosine similarity or inner product in the shared embedding space. The improvement is real. Dense representations capture semantic similarity that BM25 misses: "automobile" and "car" embed nearby even though they share no characters. But the architecture of the score is unchanged. It is a single number, positive, reflecting how close the document is to the query. It says nothing about how the document might *fail* to answer the query.

Lewis et al. (2020) formalised this into the retrieve-then-generate pipeline that now defines Retrieval-Augmented Generation (RAG). A retriever selects documents by positive relevance, and a generator conditions on whatever is retrieved. The generator has no independent access to the corpus. It sees what the retriever gives it, and it trusts what it sees. Shuster et al. (2021) confirmed the value of this trust: retrieval augmentation measurably reduces hallucination in dialogue systems compared to unaugmented generation. But the

reduction is partial. The generator still cannot distinguish a relevant document from a topically similar but inapplicable one, and the retriever gives it no signal to try. Each step in this progression was locally rational: BM25 solved ad-hoc search, dense retrieval solved semantic matching, RAG solved knowledge grounding. The positive-only assumption survived each transition because it worked well enough for the tasks at hand. The gap is an emergent property of the accumulated stack, not a failure of any single component.

That gap becomes visible when the stack is deployed in domains where "most similar" is not a reliable proxy for "most useful." Tax law, medical diagnosis, financial compliance: these are fields where a wrong answer that looks right is worse than no answer at all. The trust the generator places in the retriever was a reasonable default. It is not a safe one. One sees the same assumption in each generation of the architecture.

2.2 The fusion and reranking layer

Modern RAG systems rarely use a single retriever. The standard practice is to combine signals from several retrievers (typically BM25 and a dense model) and merge the results. Under Reciprocal Rank Fusion (RRF), the dominant merge strategy, each retriever's output is converted to a rank-based score and summed, thereby discarding the magnitude information in the original relevance scores (Bruch et al., 2023).

The consequences are specific. A document that one retriever scored at 0.95 and another at 0.51 may receive the same fused rank as a document scored 0.80 by both. The signal that one retriever was nearly certain while the other was barely above chance does not survive the merge. After fusion, the system knows which documents ranked highest. It does not know how confident it should be about any of them.

Cross-encoder reranking is the standard next step. A cross-encoder processes each query–document pair jointly through a transformer, producing a more nuanced relevance score than a bi-encoder can. The gains are well documented. But a cross-encoder is *stricto sensu* a text-relevance scorer. It reads the query text and the document text. It does not read metadata. It does not know which jurisdiction a statute belongs to, which version of a regulation it is reading, or whether the entity named in the document is the entity the query asks about. A domain-specific cross-encoder can, in principle, learn textual correlates of these signals. A reference to "Directive 2006/112/EC" carries jurisdiction and vintage in the text itself. But such learning is implicit, fragile, and unauditable. It depends on surface patterns that may or may not generalise. When the distinction lives outside the text entirely (a superseded regulation whose prose is identical to its replacement save one clause, or two provisions from different jurisdictions that use the same statutory language), the textual correlate vanishes. The cross-encoder falls back on surface similarity and has nothing better to work with. A reranker trained on general relevance judgments will not learn that "Federal Tax Act" in a Swiss provision and "Bilateral Tax Treaty" in a Swiss–German agreement are categorically different legal instruments, despite sharing the words "tax" and "Swiss."

The reranker scores text. It does not verify scope.

2.3 Three questions

The pipeline described above (retrieve by positive similarity, fuse by rank, rerank by text relevance, generate from the result) works when one condition holds: that textual relevance and practical applicability are the same thing. When they diverge, the system has no way to detect the divergence. Three questions test whether this condition holds in domains where wrong answers carry consequences.

First: if two documents are equally similar to a query but one applies to the wrong jurisdiction, what signal does the retriever have to distinguish them? The embedding captures "tax treaty residency rules." It does not capture "applicable under the Swiss–Germany bilateral agreement, not under Swiss domestic law." Both documents are about tax residency. One answers the question. The other does not.

Second: if the retriever assigns a score of 0.87 to a document, what does that number mean? Is it a probability that the document is relevant? A relative ranking within this result set? Can one threshold on it? The number is produced by a function trained to rank, not to estimate probabilities. It has the form of a confidence score without the substance of one.

Third: if a regulation was amended last year, and the current and superseded versions share 95% of their text, how does the retriever distinguish current from superseded? The embedding is computed from the text. The text is nearly identical. The legal status is opposite. No scoring function that operates on text alone can see the difference.

These are not, in our view, contrived scenarios. They describe the normal operating conditions of legal, medical, and financial retrieval. A tax lawyer asking about a bilateral treaty provision will face jurisdictional ambiguity. A clinician querying drug interactions will face version-sensitive guidelines. A compliance officer verifying regulatory requirements will need to know whether the retriever's confidence is meaningful. Kanhabua et al. (2015) surveyed two decades of temporal information retrieval research and catalogued sophisticated solutions for time-aware search, solutions that have not been integrated into standard RAG architectures. The temporal dimension is one example. Jurisdiction and calibration are two others.

The next section presents evidence that these questions have measurable, documented, and in some cases dangerous answers.

3 The case against positive-only retrieval

Consider a RAG system answering the Swiss–Germany DTA query from Section 1. The retriever returns ten chunks ranked by cosine similarity. Eight are about tax treaty residency rules. Two are about Swiss domestic tax law, wrong jurisdiction, near-identical vocabulary. The system has no representation of "this document does not apply to the bilateral treaty context." It knows only that the document is similar. Evidence from retrieval evaluation, calibration theory, and temporal information retrieval converges on the same conclusion: positive-only relevance scoring fails in systematic, predictable, and dangerous ways (Table 1 summarises the evidence across domains; Figure 2 illustrates the embedding-space problem).

3.1 Similarity is not applicability

Reuter et al. (2025) identify a failure mode they call *Document-Level Retrieval Mismatch* (DRM): the proportion of top-k retrieved chunks originating from the wrong source document. In legal corpora, structurally similar contracts and statutory provisions embed as near-identical vectors. The retriever surfaces the right topic from the wrong source, namely a clause from a different contract that happens to use the same phrasing. For legal applications this is not a retrieval quality problem. It is a validity problem. A contract clause pulled from the wrong agreement has zero legal force, regardless of how textually similar it is.

Zheng et al. (2025) confirm this at the statutory level. Their Housing Statute QA benchmark contains housing and eviction provisions from over 50 US jurisdictions. BM25 and dense retrievers cannot distinguish the correct jurisdiction from the wrong one, because the texts are topically identical and differ only in which state enacted them. The embedding space represents "this statute is about eviction." It does not represent "this statute is from California, and the question asks about New York." That is a different question.

Hindi et al. (2025) survey the legal RAG landscape more broadly and confirm that retrieval precision remains the primary bottleneck across implementations; the problem is systemic, and is not confined to any single architecture or dataset.

The pattern is not limited to law. Zhao et al. (2025) document the same failure in medical retrieval, where diseases with overlapping symptoms cause standard RAG systems to surface records for the wrong diagnosis. Their MedRAG system addresses this by building a hierarchical diagnostic knowledge graph that encodes the negative signal "these two conditions share symptoms but are categorically different." The underlying problem is severe: an estimated 795,000 Americans annually suffer permanent disability or death from misdiagnosis of confusable diseases (Newman-Toker et al., 2023).

In financial services, the false-positive rates are starker still. Daivam (2025) reports that a RAG semantic caching system in banking produced false-positive rates exceeding 99% for some embedding models at a standard 0.7 similarity threshold. A customer asking to cancel a credit card was matched to investment account closure procedures with 88.7% confidence. The higher the similarity score, the more confidently the system delivered the wrong answer.

A Microsoft Health and Life Sciences technical report (2025) demonstrated why this happens at the metric level. They tested whether standard NLP similarity measures could distinguish clinically equivalent statements from clinically opposite ones. They could not. BERTScore, ClinicalBERT, and MoverScore all scored the opposite-meaning statement *higher* than the correct equivalent. Domain-specific embeddings improved differentiation by roughly 5%. Nowhere near sufficient.

Barnett et al. (2024) call this *silent wrong retrieval*, and state it is the most dangerous of their seven catalogued RAG failure points, because the generation model cannot detect it. The retrieved text *looks* right. It is noise dressed as signal. The tractability of this problem is well-established. Qu et al. (2021) introduced denoised hard negative mining for dense passage retrieval, explicitly filtering documents that resemble the query but fail to answer

it. Meghwani et al. (2025) extend the approach to domain-specific enterprise retrieval, yielding 15% improvement in Mean Reciprocal Rank (MRR@3) by mining contextually irrelevant near-misses. The mechanism works. It is simply absent from standard RAG architectures.

The problem has a further dimension: retrieved documents can actively conflict with each other or with the model's parametric knowledge. Feldman et al. (2024) show that retrieval augmentation can paradoxically *increase* hallucination when the retrieved context contradicts what the model already knows, the "double-edged sword" of RAG. Zhang et al. (2025) formalise this as the knowledge conflict problem and show that current systems have no principled mechanism for resolving it: their KnowPO framework addresses knowledge selection at the generation layer, but the conflict signal is not represented at retrieval time, where it could inform the system that the evidence is contested rather than consistent. A retrieval layer that typed its negative evidence could distinguish "I found strong support" from "I found two sources that contradict each other", the difference between CRANE's Pattern A and Pattern B.

Were the system to check for scope mismatches (e.g. entity, jurisdiction, application context) at retrieval time, these false positives would be caught before reaching the generator. Under standard similarity-based retrieval, they are invisible.

The representational diagnosis is precise: positive-only retrieval encodes what a document *is about* but has no signal for what it *does not apply to*. A document can be topically relevant and jurisdictionally inapplicable. Semantically similar and factually wrong. The system needs a way to say "right topic, wrong scope," and currently it cannot.

3.2 Confidence without calibration

A retrieval score of 0.87 tells you that document d is in the 87^{th} percentile of similarity to query q in the current result set. It does not tell you that d has an 87% chance of being relevant. The score says nothing about whether d actually answers q.

This distinction matters because downstream components (the reranker, the generation prompt, any confidence threshold) treat the score as if it were a probability. It is not. Guo et al. (2017) demonstrated that modern neural networks are systematically miscalibrated: a model predicting with 90% confidence is correct far less than 90% of the time. They introduced Expected Calibration Error (ECE) and showed that temperature scaling, a single learned parameter, substantially reduces the gap. The finding reshaped how the machine learning community thinks about classifier confidence. It has not yet reshaped how the retrieval community thinks about relevance scores.

Penha and Hauff (2021) ground the problem in the Probability Ranking Principle (PRP), the foundational IR result stating that ranking documents by $P(\text{relevant} \mid d, q)$ is optimal. The PRP requires two conditions: the model must be well-calibrated (C1) and must report confidence with certainty (C2). BERT-based rankers satisfy neither. Thus the theoretical guarantee that underpins modern neural ranking, that ranking by score is ranking by relevance probability, does not hold in practice.

The consequences are not theoretical. Omar et al. (2025) report a statistically significant inverse correlation between model confidence and accuracy ($r = -0.40$, $p = .001$) across 1,965 medical questions tested on 12 LLMs. The worst-performing models exhibited the highest confidence. Ozaki et al. (2024) find that RAG itself can make this worse: inserting retrieved documents improves answer accuracy but produces unwarranted certainty, because the model cannot distinguish subtly wrong context from genuinely supportive context.

The word "relevant" means different things depending on who uses it. A retrieval model means "textually similar." A practitioner means "actually answers my question under the applicable legal framework." They mean different things by it.

This is in our view the central problem with positive-only retrieval. Without calibration, a RAG system cannot distinguish "I found a strong answer" from "I found something that looks like an answer." In domains where the cost of a confident wrong answer vastly exceeds the cost of admitting uncertainty, that distinction is not optional. The machine learning community has known since Brier (1950) and Platt (1999) that scores can be post-hoc calibrated, and since Niculescu-Mizil and Caruana (2005) that different model architectures carry different calibration properties. The retrieval community has, with notable exceptions (Cohen et al., 2021; Yan et al., 2022), largely ignored this work.

The representational diagnosis: the system produces a ranking but not a calibrated probability. It can order documents but cannot say how confident it is, in any meaningful sense, that a given document answers the question. Without $P(\text{relevant})$, no downstream component, whether reranker, generator, or confidence gate, can make a principled decision about how to respond.

3.3 Temporal blindness

Regulations are amended. Statutes are repealed. Clinical guidelines are superseded. In each case, the current and former versions share most of their text. A dense retriever encodes both as near-identical vectors, because the text *is* near-identical. What changed is the legal status and applicability. Not the text.

De Martim (2025) puts it directly: standard RAG is "temporally naive." SAT-Graph RAG models each version of a legal structural component as a distinct node in a knowledge graph, with temporal states and legislative events as first-class entities. Applied to the Brazilian Constitution, the system demonstrates that flat-text retrieval is blind to the diachronic structure of law. A provision valid in 2018 and repealed in 2021 embeds identically to the current version, thereby making wrong-version retrieval a systematic rather than accidental failure.

Huwiler et al. (2025) quantify the gap: standard RAG achieves only 58–64% accuracy on version-sensitive queries. Their VersionRAG system routes queries through specialised paths: version queries use graph traversal, content queries use vector search. The baseline numbers are the finding that matters. A system that is wrong 36–42% of the time on questions about which version of a document is current is not a system one can deploy in

a regulated environment. One could argue that version management is not only a retrieval problem, but also a metadata problem, not a retrieval problem. It is both.

The temporal dimension compounds the other two failures. Seeing that the pre-2024 and post-2024 versions of the Swiss–Germany DTA share roughly 95% of their text, the retriever assigns near-identical similarity scores to both. The post-2024 version addresses remote workers; the pre-2024 version does not. No confidence score derived from text similarity alone can distinguish them, because the distinction lives in the metadata, not in the prose. *A fortiori*, a system that lacks calibrated confidence (Section 3.2) and cannot represent scope mismatches (Section 3.1) has no chance of handling temporal shifts correctly.

This is not a new problem for information retrieval *stricto sensu*. Berberich et al. (2007) solved point-in-time text search two decades ago with time-stamped inverted indices. Gade et al. (2024) show that a lightweight temporal score (the reciprocal of the time difference between query and document timestamps) yields 74–165% improvement in recall, requiring no retraining. The tools exist. They have not been integrated into RAG. The reason is architectural: classical temporal IR operates on whole documents with structured metadata, while RAG operates on chunks stripped of document-level context. A chunk of a repealed regulation carries no timestamp, no version identifier, no link to its successor. The temporal signal is lost at ingestion, and no amount of retrieval-time scoring can recover what was discarded.

The temporal negative is equally familiar in software engineering. Zhou and Walker (2016) find that API deprecation does not follow a clean lifecycle: functions are deprecated, un-deprecated, and re-deprecated without predictable sequence. The parallel to legal norm evolution is exact. A provision can be enacted, repealed, and re-enacted, and at each transition the text may be identical while the legal status reverses. No system relying on text similarity alone can track this.

The representational diagnosis mirrors Section 3.1: the system encodes what the document says but not *when* it says it. A document that was relevant in 2022 and is no longer relevant in 2025 carries no signal marking the change. The system needs a way to say "this version is superseded," a temporal negative, and it has none.

We have drawn evidence from domains where two conditions hold: high semantic similarity between correct and incorrect answers is structurally inherent (not incidental), and confident wrong answers carry asymmetric costs. Legal provisions across jurisdictions, drug interactions across conditions, financial products across regulatory regimes. These are not cherry-picked failure cases. They are domains where the architecture of the problem guarantees that positive-only retrieval will fail. In low-stakes retrieval (web search, entertainment recommendations, casual question answering), the failure modes documented above are tolerable. A wrong movie recommendation is not a wrong legal opinion. The argument in this paper is scoped accordingly.

One might ask whether these failures could be fixed downstream: by a better reranker, a verification agent, or a more careful generation prompt. They cannot, for a structural reason: by the time any downstream component sees the retrieved documents, the wrong document is already in the candidate set, and it looks indistinguishable from the right one.

A cross-encoder reranker is a text-relevance scorer; it cannot penalise a document for being from the wrong jurisdiction unless that signal is encoded before reranking. A verification agent can fact-check generated claims, but it cannot detect that the source document, while factually accurate *in its own context*, does not apply to the query's context. The fix must be at retrieval time, because that is where the scope, confidence, and temporal signals either exist or do not.

These three failure modes are not independent. They share a common cause: retrieval systems model what documents *are* about but have no representation of what documents are *not* about. Each strand identifies a specific signal that is missing (scope applicability, calibrated confidence, temporal validity), and in each case the missing signal corresponds to a type of negative evidence that the system cannot express. The next section derives the properties that any adequate system must satisfy and presents one architecture meeting all three.

4 Crane: proof by construction

The three representational gaps identified in Section 3 (missing scope signal, missing calibrated confidence, missing temporal validity) are not independent engineering problems. They are symptoms of a single architectural omission: retrieval systems encode what a document is about but have no mechanism for encoding what it is not about. Any system adequate for high-stakes domains must address all three. We derive the required properties below, then present one architecture that satisfies them (Figure 4 maps the full derivation from evidence to mechanisms).

4.1 Required properties

Each property corresponds to a representational gap documented in Section 3. The derivation is meant to be minimal: we state what a system *must* do, namely satisfy three functional requirements, not how it should do it.

> **Property 1 (Negative evidence representation).** For each document–query pair, the retrieval system must represent not only relevance but also specific ways in which the document fails to answer the query, typed by failure mode.

This follows from Section 3.1. A retriever that assigns a single positive relevance score to a document cannot distinguish "highly relevant" from "topically similar but jurisdictionally inapplicable." The negative signal must be typed because different failure modes require different responses: a scope mismatch (wrong entity, wrong jurisdiction) calls for exclusion; a sibling match (related but distinct concept) calls for disambiguation; a temporal mismatch (superseded version) calls for version resolution. A single "not relevant" flag is insufficient. The failure mode matters.

Property 2 (Calibrated confidence). The retrieval system must produce, for each document–query pair, a calibrated relevance probability $P(\text{relevant}) \in [0, 1]$, not a ranking score.

This follows from Section 3.2. The Probability Ranking Principle guarantees optimal ranking only when the scores are calibrated probabilities (Penha and Hauff, 2021). Current systems violate this assumption. Without $P(\text{relevant})$, no threshold is meaningful, no generation strategy can be conditioned on confidence, and no downstream component can distinguish a strong answer from a plausible-looking wrong one. The requirement is not merely that scores be normalised; it is that they be calibrated in the technical sense: among all documents assigned confidence p, approximately fraction p should be truly relevant.

Definition 1 (Calibrated retrieval). A retrieval system is *calibrated* if

$$P(\text{relevant} \mid \text{confidence}(d, q) = p) \approx p \quad \text{for all } p \in [0, 1].$$

This definition extends the standard classification calibration criterion (Guo et al., 2017) to document–query pairs. It is a functional requirement, not a metric proposal. To our knowledge, no existing RAG system satisfies it.

Property 3 (Confidence-gated generation). The generation layer must receive not a ranked list of documents but a *confidence profile* (including the type and severity of any negative evidence) and must select its generation strategy accordingly.

This follows from the interaction of all three gaps. A generator that receives a ranked list treats all items equally: it synthesises from whatever it is given. But a list containing a scope negative (wrong jurisdiction) requires a different response from a list containing a temporal negative (superseded version) or a list where confidence is genuinely low. The generation strategy must thus be conditioned on the *type* of uncertainty, not just its magnitude. Assert, hedge, flag a boundary, acknowledge a gap. These are different responses to different situations, and a system that cannot distinguish them will default to asserting, which is the most dangerous option.

4.2 A taxonomy of negative evidence

We propose a three-type classification of negative retrieval evidence (Figure 3). Each type corresponds to a failure mode documented in Section 3 and to a distinct required response.

Definition 2 (Negative evidence taxonomy).

Scope negatives. The document addresses the correct topic but the wrong entity, jurisdiction, or application context. The document is *about* the query's subject but does *not apply* to the query's scope. *In casu*: a Swiss domestic tax provision retrieved for a bilateral treaty question.

Sibling negatives. The document addresses a related but distinct concept that shares vocabulary with the query's target. The document is *near* the correct answer in embedding space but *categorically different* in domain semantics. *In casu*: a provision on permanent establishment retrieved for a question about residency, because both fall under "tax treaty" and share key terms.

Temporal negatives. The document was applicable at a different point in time than the query requires. The document *was* relevant but *is not currently* relevant, or vice versa. Temporal negatives operate at claim-level granularity: individual provisions within a single document may have different temporal statuses (e.g., Article 15(1) unchanged while Article 15(2) was amended), and the system must track applicability per claim, not per document. *In casu*: the pre-2024 version of the Swiss–Germany DTA retrieved for a question about post-2024 remote workers.

The taxonomy is not claimed to be exhaustive. Other types may emerge from empirical study, *de lege ferenda*: jurisdictional hierarchy negatives (applicable law but wrong court level), for instance, or conditional negatives (applicable only if a factual predicate holds). What we claim is that these three types cover the failure modes most extensively documented in the current literature, and that any system addressing them would represent a structural improvement over positive-only retrieval.

4.3 CRANE instantiation

CRANE (Calibrated Retrieval with Adversarial Negative Evidence) is one architecture satisfying Properties 1–3. We specify it at the level of functional mechanisms, what each component must do, what inputs it receives, and what outputs it produces, demonstrating that the three properties can be jointly satisfied in a single retrieval pipeline. Other architectures satisfying the same properties would share the same structural advantages.

Mechanism 1: Typed negative evidence at retrieval time

CRANE operates on *natural semantic units*, segments that preserve the internal coherence of a legal provision, clinical guideline, or regulatory clause, rather than arbitrary fixed-length chunks. For each candidate unit, the retrieval layer checks for scope, sibling, and temporal negatives in addition to computing positive relevance. Scope negatives are detected through metadata comparison: does the document's jurisdictional scope, entity type, or application context match the query's? Sibling negatives are detected through semantic disambiguation: is the concept addressed by the document the *same* concept the query asks about, or a related but distinct one? Temporal negatives are detected through version metadata: is this document the current version, or has it been amended, repealed, or super-

seded? The negative evidence is typed and attached to the document's retrieval record. It is not discarded. The technical effect is a structured signal, attached to the retrieval record at indexing time, that enables downstream components to distinguish topically relevant from contextually inapplicable documents before generation.

Mechanism 2: Calibrated merge scoring

Multiple retrieval signals (positive relevance from one or more retrievers, negative evidence by type, structured metadata features) are combined by a pluggable merge scorer. The scorer produces $P(\text{relevant}) \in [0, 1]$: a calibrated probability satisfying Definition 1, not a ranking score. A cold-start implementation can use weighted combination and produce useful results immediately. A trained implementation can use a learned scoring model to map features to calibrated probabilities, trained on relevance judgments, click data, or expert annotations. The scorer operates on a *relevance graph* that captures not only document–query relationships but also inter-document relationships: one provision superseding another, two provisions addressing the same subject from different jurisdictions, a regulation and its implementing directive. The graph structure enables negative-evidence propagation; a temporal negative on one version propagates to related versions through the graph. The technical effect is a probability estimate with a defined calibration property (Definition 1), enabling threshold-based decisions that ranking scores cannot support.

Mechanism 3: Confidence-gated generation

The generation layer receives a confidence profile per retrieved document cluster, not a flat ranked list. The profile includes the calibrated relevance score, the types of negative evidence detected, and a confidence pattern classification. The technical effect is response differentiation: the system's output changes not only in content but in epistemic posture based on the type and severity of uncertainty in the evidence. Five patterns determine the generation strategy (Table 2):

> **Pattern A (Clean).** High confidence, no negatives detected. The system asserts with direct citation, satisfying the Attributable to Identified Sources criterion (Rashkin et al., 2022).

> **Pattern B (Contested).** Multiple sources, conflicting signals. The system presents competing positions and marks the boundary.

> **Pattern C (Confused).** Low confidence, sibling negatives dominant. Hedged response; the system says it is not certain which concept applies.

> **Pattern D (Temporal Boundary).** Temporal negatives present. The system flags version dependencies explicitly: "this answer applies as from [date]; the prior version does not address [topic]."

> **Pattern E (Sparse).** Insufficient evidence. No confabulation. The system acknowledges the gap rather than filling it with plausible-sounding noise.

Each generation cycle produces a *claims artifact*: a structured record of what the system asserted, under which confidence pattern, against which evidence base (with constraint signatures), and which negative evidence was considered and how it affected the response. The claims artifact is the audit trail that makes Crane's reasoning inspectable and reproducible, a requirement in regulated domains where the basis for a system's output must be traceable. Before presentation, a deterministic validation step verifies that the generated response is consistent with the confidence profile: a Pattern A assertion that cites a document flagged with temporal negatives fails validation and is regenerated under the correct pattern. The validation operates on the claims artifact, not on the generated text, ensuring consistency between what the system decided and what it said.

Returning to the running example. The Swiss–Germany DTA query triggers Crane's scope check, which detects that two of the ten retrieved documents are Swiss domestic provisions, not bilateral treaty provisions. The temporal check flags that one retrieved chunk is from the pre-2024 treaty version. The merge scorer integrates these signals and produces calibrated confidence estimates. The generator receives a Pattern D profile and responds accordingly: it cites the post-2024 bilateral provision, flags the temporal boundary, and notes that the pre-2024 version does not address remote workers.

A further consequence of typed negative evidence is constraint-signature-aware deduplication. Each retrieved document carries a *constraint signature*: the combination of negative-evidence types and severity levels attached by the detection layer. Documents with identical constraint signatures (same scope status, same temporal status, same sibling classification) can be grouped and deduplicated before scoring, reducing merge scorer computation without information loss. Documents with different constraint signatures, even if textually similar, are preserved as distinct evidence, because their failure modes differ. Standard deduplication operates on text similarity or embedding proximity and cannot make this distinction.

Several existing systems address fragments of this architecture. Hard negative mining (Qu et al., 2021; Meghwani et al., 2025) identifies documents that *look* relevant but are not, yet produces an undifferentiated negative signal, not a typed one. VersionRAG (Huwiler et al., 2025) and SAT-Graph RAG (De Martim, 2025) solve temporal version resolution but do not produce calibrated confidence or gate generation strategy. Self-RAG (Asai et al., 2024) evaluates retrieval quality at generation time but makes binary useful/not-useful judgments without typing the failure mode. Corrective RAG (Yan et al., 2024) filters retrieved documents but does not feed the filter's signal into generation strategy selection. FLARE (Jiang et al., 2023) and DRAGIN (Su et al., 2024) use generation-time confidence to trigger additional retrieval when the model is uncertain, a binary retrieve-or-not decision, not a five-pattern typed classification of uncertainty type. KnowPO (Zhang et al., 2025) resolves knowledge conflicts via preference optimisation at the generation layer, not at retrieval time. The closest precedent for the merge scorer itself is McMahan et al. (2013), who deploy gradient-boosted decision trees (the ensemble method formalised by

Friedman, 2001), producing calibrated click probabilities under zero-latency constraints, with validation gating and auto-rollback, for ad-click prediction at scale. Covington et al. (2016) apply a similar calibrated-probability architecture to YouTube recommendation ranking. The architectural parallel is real: all three systems use GBDT-class or deep scoring models to produce calibrated probabilities from heterogeneous features. The distinction is the feature space and the downstream consumer. Crane's scorer ingests negative-evidence activation signals (scope, sibling, temporal flags per document) that have no analogue in ad-click prediction, and its calibrated output feeds a confidence classifier and tier router that selects among five generation strategies, a downstream architecture absent from McMahan's system, where the probability feeds a single auction mechanism. What is absent from all of these systems is the feedback loop: first, typed negative evidence is detected at retrieval, this is then integrated into calibrated scoring and then finally used to select among differentiated generation strategies. Each existing system solves one link in the chain. Crane's contribution is the chain itself.

The mechanisms above specify what a Crane implementation must do at each stage: detect typed negatives via metadata and semantic checks, merge signals into calibrated probabilities, and gate generation on confidence profiles. Companion technical documentation details specific detection algorithms, scorer feature sets, and threshold configurations; what this paper establishes is the functional architecture, the information flow from typed negative detection through calibrated scoring to differentiated generation, and the properties it must satisfy. A system implementing these mechanisms satisfies Properties 1–3 and thereby addresses the three failure modes documented in Section 3.

Four architectural choices in Crane are not arbitrary and their alternatives are inferior for high-stakes domains. *First*, negative evidence is represented as structured metadata, not as embedding-space signals. Embedding-space negatives are opaque (the model cannot explain *why* a document fails), non-compositional (combining scope and temporal negatives requires ad hoc operations in continuous space), and subject to the embedding dimension ceiling identified by Weller et al. (2026). Structured metadata supports typed per-failure-mode penalty weights that are interpretable and auditable. *Second*, validation is deterministic, not LLM-based. An LLM-based validator produces non-deterministic, non-reproducible, non-auditable outputs, unacceptable in legal and medical domains where the same query must yield the same validation result on repeated execution. Crane's scope check is a metadata comparison; its output is deterministic and logged. *Third*, the architecture requires zero new lookups at query time beyond the initial retrieval. Typed negative evidence is attached at ingestion; the merge scorer operates on the retrieval record. This is not merely a latency optimisation; it is a system-architectural invariant that produces cascading deterministic guarantees from scorer through classifier to router. *Fourth*, the self-improving loop (Section 6) operates on individual artifacts (detection rules, scorer weights, generation templates), not on model retraining. This is categorically different from MLOps: individual artifact rollback rather than whole-model rollback, per-artifact staging rather than model-level canary deployment, deterministic scope rather than the Changing Anything Changes Everything problem identified by Sculley et al. (2015).

4.4 Formal properties

Two formal observations connect CRANE to the calibration literature.

First, calibrated retrieval (Definition 1) is a stricter requirement than ranking quality. A system can achieve high normalised discounted cumulative gain (nDCG) or MRR while producing scores that are not calibrated: the ordering is correct but the magnitudes are meaningless. For generation gating, the magnitudes matter. Pattern A requires not just that the top-ranked document is relevant, but that the system's confidence in its relevance is justified. Elkan (2001) showed that in domains where different types of misclassification carry different costs, calibrated probabilities are necessary for optimal decision-making. Legal retrieval is precisely such a domain: a confidently wrong answer (false positive) is far more costly than a missed answer (false negative). Lin et al. (2017) formalised the asymmetric loss principle with focal loss, which down-weights easy classifications to focus on hard, ambiguous cases, the cases where negative evidence matters most.

Second, the negative evidence taxonomy operates in a representational space that is not constrained by the embedding dimension ceiling identified by Weller et al. (2026). Their proof shows that the number of distinct top-k retrieval results expressible by a d-dimensional embedding is fundamentally bounded. Typed negative evidence is not an embedding-space signal. It is a metadata-level and semantic-level check that operates orthogonally to the embedding, whereby the system consults structured attributes rather than geometric proximity. A scope negative is detected by comparing jurisdiction fields, not by measuring cosine distance. This means the representational capacity for negative evidence scales with the metadata schema, not with the embedding dimension. The ceiling that limits positive-only retrieval does not apply.

In our view, this is the strongest theoretical argument for negative evidence as a *complement* to embedding-based retrieval rather than a replacement for it. Better embeddings improve positive relevance estimation. They cannot, by construction, provide the typed negative signals that high-stakes domains require. The two work in different representational spaces and address different failure modes. A system needs both.

5 Potential objections

The argument in Sections 3 and 4 invites four objections. We state each in its strongest form and respond.

5.1 "Better embeddings will solve this"

The strongest version of this objection runs as follows. Embedding models are improving rapidly. Fine-tuned domain-specific embeddings, larger vector dimensions, and better training data will eventually capture the distinctions that current models miss. The scope, sibling, and temporal failures documented in Section 3 are transient, artifacts of immature models, not of architectural limitation.

The evidence points the other direction. Weller et al. (2026) prove that the number of distinct top-k retrieval results expressible by a d-dimensional embedding is fundamentally

bounded. This is a mathematical ceiling on any single-vector representation, not an artifact of current models. On their LIMIT benchmark, state-of-the-art embedders fail on trivially simple queries where documents differ by a single attribute, queries any human would answer correctly. Better training cannot overcome a geometric constraint.

Sciavolino et al. (2021) confirm this at the empirical level. Their EntityQuestions benchmark tests queries whose correct answer depends on which entity is named, not which topic is discussed. Dense retrievers drastically underperform BM25. The distinction is structural (which entity?) rather than semantic (what topic?). Dense retrieval's advantage (capturing semantic similarity) is what makes it worse at scope-level discrimination: it groups documents by topic, not by the entity or jurisdiction they apply to.

Multi-vector and late-interaction models (Khattab and Zaharia, 2020) bypass the single-vector ceiling by producing per-token embeddings and matching at the token level. The improvement in granularity is real. But token-level matching still operates in text-semantic space. A ColBERT model that correctly matches the tokens "Article 15(2)" and "cross-border workers" will give high scores to both the correct bilateral treaty provision and the wrong Swiss-domestic provision, because the discriminating information (which jurisdiction?) is not in the tokens being matched. Finer granularity surfaces more plausible-looking wrong documents, not fewer.

The point generalises. The more aggressively a model groups semantically similar documents, the harder it becomes to distinguish applicable from inapplicable within that group. Yousuf et al. (2026) demonstrate this directly: the disambiguating signals live in structured metadata (jurisdiction, date, entity type), not in text semantics. No embedding improvement, single-vector or multi-vector, captures what is not in the text.

5.2 "Post-hoc reranking fixes this"

Grant that the retriever is blind to scope and calibration. A cross-encoder reranker jointly processes query–document pairs and can learn arbitrary relevance functions. A well-trained reranker could learn to detect wrong-scope or wrong-jurisdiction documents. More recent work goes further: Self-RAG (Asai et al., 2024) and Corrective RAG (Yan et al., 2024) equip the generator itself with the ability to evaluate and filter retrieved evidence.

Two problems. *First*, cross-encoders and self-reflective generators operate on the retriever's candidate set. If the top-k already contains the wrong documents (and Section 3.1 showed it systematically does), then reranking and self-reflection rearrange or filter within a contaminated pool. Wang et al. (2021) showed that dense retrievers need BM25 interpolation precisely because they miss relevance signals that keyword matching catches. The reranker inherits whatever the retriever produces.

Second, the computational economics do not scale. Weller et al. (2026) report that a cross-encoder (Gemini-2.5-Pro) solves 100% of their LIMIT benchmark queries, but at the cost of processing every document in the corpus as a query–document pair. For any corpus of meaningful size, this is infeasible. The retriever must thus do the filtering.

Self-RAG deserves separate consideration because it is a genuine architectural intervention, not a post-hoc patch. It trains reflection tokens into the model, making retrieval

evaluation a first-class operation. The question is what kind of evaluation it performs. Self-RAG makes a binary judgment: useful or not. It cannot tell the generator *why* a document fails or how to adjust the response accordingly. CRANE provides typed failure modes (scope, sibling, temporal) and calibrated confidence, enabling differentiated generation strategies (Patterns A through E). A system that can say "this document is from the wrong jurisdiction" produces a different response from one that can only say "this document seems unhelpful." The distinction is operational. Self-RAG evaluates retrieval quality at inference time, per query, a computational cost that scales with query volume. CRANE attaches typed negative evidence at indexing time (once per document, as structured metadata) and evaluates against it at query time via the merge scorer. The scope check is a metadata comparison, not a neural inference. The technical effect is different: Self-RAG improves generation quality by filtering bad retrievals after they occur; CRANE prevents bad retrievals from reaching the generator by encoding the failure signal in the retrieval record itself.

HyPA-RAG (Kalra et al., 2024) takes a complementary approach for legal applications, combining dense retrieval, sparse retrieval, and knowledge graph methods with adaptive parameter tuning based on query complexity. The system improves retrieval quality for policy questions but does not produce typed negative evidence or calibrated confidence scores. It selects among retrieval strategies; it does not characterise *why* a retrieved document might be inapplicable. The distinction is architectural: HyPA-RAG optimises *which* retrieval method to use for a given query, while CRANE characterises *what the retrieved documents do not answer* regardless of how they were retrieved. KnowPO (Zhang et al., 2025) operates at yet another layer, optimising the generator's response to knowledge conflicts through preference learning. None of these systems produces the feedback loop that CRANE requires: typed negative evidence at retrieval → calibrated merge scoring → confidence-gated generation. Each addresses a fragment of the problem; CRANE's contribution is the information architecture connecting them.

5.3 "This is too complex to deploy"

Adding negative evidence checks, calibrated scoring, and confidence-gated generation to a RAG pipeline sounds like significant engineering effort. Simpler alternatives exist: larger context windows that can absorb more documents and let the model sort them out, or better prompting that instructs the generator to be cautious.

Neither addresses the core problem. A larger context window does not help the model distinguish applicable from inapplicable documents; it merely ensures both are present. Were one to place ten documents in the context window, three of them scope negatives, the generator still has no signal that those three do not apply. It will synthesise from all ten. Liu et al. (2024) showed that LLM attention degrades for information positioned in the middle of long contexts; a million-token window gives capacity but not discrimination. More fundamentally, a long-context approach makes jurisdictional and temporal reasoning *possible* but not *mandatory*. There is no mechanism forcing the model to check scope before synthesising. CRANE makes the check mandatory and the failure mode visible. A

cautious prompt ("only cite documents you are confident about") provides no mechanism for confidence; it adds a hedging instruction on top of uncalibrated scores.

Individual techniques address fragments of the problem. Hard negative mining improves retrieval quality (Meghwani et al., 2025). Metadata-aware indexing enriches chunks with structured fields (Yousuf et al., 2026). Temporal scoring improves recall on time-sensitive queries, requiring no retraining (Gade et al., 2024). But none of these techniques, applied independently, produces typed negative evidence that feeds a calibrated merge scorer that conditions generation strategy. The integration is where the system-level properties emerge: a hard negative miner does not produce calibrated $P(\text{relevant})$, a temporal scorer does not distinguish scope negatives from sibling negatives, and a metadata index does not gate generation. CRANE's contribution is the information architecture connecting these signals so that each enables the next.

5.4 "Where is the empirical proof?"

The evidence standard for this paper is evidence synthesis and proof by construction, not controlled experiment. That is a deliberate methodological choice, not a gap. Section 3 synthesises evidence from legal, medical, and financial retrieval showing that positive-only scoring fails in systematic, documented ways. Section 4 demonstrates, by construction, that the three required properties (typed negative evidence, calibrated confidence, confidence-gated generation) are jointly satisfiable in a single retrieval pipeline. Section 6 states four falsifiable predictions to guide empirical evaluation. The methodological question is whether the problem characterisation and the architectural derivation are sound, and that question is answerable from the evidence presented.

The individual mechanisms draw on established techniques (metadata comparison for scope checking, learned scoring models for calibration, template-conditioned generation for response differentiation), each with documented technical effects on retrieval precision, calibration error, and generation quality in the literature cited. CRANE's contribution is the feedback architecture connecting them: typed negative evidence at retrieval feeds calibrated merge scoring, which conditions generation strategy. No existing system closes this loop. The architectural integration produces a technical effect, measurable reduction in false-positive retrieval rates and improved calibration, that no individual component achieves alone.

Precedent supports the approach. Sculley et al. (2015) reshaped how the field thinks about technical debt in ML systems through analysis and architectural argument. Bender et al. (2021) changed the conversation about large language models through problem identification and evidence synthesis. Both earned their impact by naming problems the field had not yet named and proposing vocabulary to reason about them. Whether CRANE is the optimal architecture for satisfying Properties 1–3 is an empirical question. Whether RAG systems for high-stakes domains need these properties is the argument this paper makes, and the evidence in Section 3 is the basis for it.

6 Implications and research roadmap

6.1 Falsifiable claims

The argument presented in this paper makes predictions that any research group can test. We state four.

First, a retrieval system implementing typed negative evidence (scope, sibling, temporal) will produce fewer false-positive retrievals on cross-jurisdiction legal queries than an equivalent system using positive-only relevance scoring, measured by Document-Level Retrieval Mismatch (Reuter et al., 2025).

Second, a merge scorer producing calibrated $P(\text{relevant})$ will enable meaningful confidence thresholds. Binning documents by assigned confidence and measuring observed relevance rates should show calibration (ECE < 0.10) that positive-only scorers cannot achieve.

Third, a generation system gated on confidence profiles will produce fewer assertive responses to queries where the evidence is weak or contested, compared to a system receiving a flat ranked list.

Fourth, temporal negative detection will reduce wrong-version retrieval rates on versioned legal and regulatory corpora, measurable against VersionRAG-style benchmarks (Huwiler et al., 2025).

These predictions are specific enough to be falsified. If a positive-only system achieves equivalent false-positive rates, calibration, generation quality, and temporal accuracy, the argument in this paper is wrong. We would welcome that result.

Two architectural extensions follow naturally from Crane's design but extend beyond the scope of this paper, which establishes the core three-property architecture.

Tiered retrieval. The mechanisms described in Section 4 assume a single retrieval stage. In practice, different query types may benefit from different retrieval strategies: exact-match for provision identifiers, semantic search for conceptual queries, graph traversal for relational queries ("what superseded this provision?"). A tiered retrieval architecture, where the query is routed to the appropriate tier and each tier produces typed negative evidence with tier-specific confidence characteristics, would extend Crane's calibration guarantees across retrieval modalities. The relevance graph introduced in Section 4.3 provides the substrate for graph-aware retrieval tiers. The specific tier definitions, routing logic, and cross-tier calibration are detailed in companion technical documentation.

Artifact-level continuous improvement. Crane's architecture separates detection rules, scorer weights, and generation templates as individually versioned artifacts. This separation enables a continuous improvement loop that does not require model retraining: detection rules are updated when new negative-evidence patterns are observed, scorer weights are recalibrated as relevance judgments accumulate, and generation templates are refined based on output quality signals. Each artifact can be staged, validated, and rolled back independently, a property that distinguishes artifact-level improvement from the Changing Anything Changes Everything problem in ML systems (Sculley et al., 2015), where chang-

ing one component's behaviour cascades unpredictably through the system. The claims artifact (Section 4.3) provides the feedback signal: each generation cycle records what was asserted and what evidence supported it, creating a closed loop from output quality back to component improvement. The specific learning paths and staging mechanisms are detailed in companion technical documentation.

6.2 Cross-domain applications

The properties derived in Section 4 are not specific to law (Table 3 maps negative evidence types across domains). They apply wherever two conditions hold: high semantic similarity between correct and incorrect answers is structurally inherent, and the cost of a confident wrong answer exceeds the cost of admitting uncertainty. Figure 5 outlines the resulting research agenda.

In medical retrieval, contraindications are scope negatives. A drug prescribed for Condition A may be contraindicated for Condition B; a document about the drug is topically relevant but clinically dangerous for the wrong patient. Chapman et al. (2001) operationalised negation detection in clinical text with NegEx; the extension to retrieval-level negative evidence is natural. Zhao et al. (2025) show the stakes: MedRAG addresses diagnostic confusion among diseases with overlapping symptoms, a problem estimated to cause 795,000 cases of permanent disability or death annually in the United States alone (Newman-Toker et al., 2023). Kim et al. (2025) find that retrieved medical content covers only 33% of must-have clinical statements. The retrieval layer is not returning enough of what matters and too much of what does not.

In financial services, jurisdictional restrictions operate identically to scope negatives. A financial product legal in Switzerland may be prohibited in the EU. Sanctions lists change weekly. FinSage (Wang et al., 2025) documents the retrieval failures, but no current system models negative jurisdictional scope. The Financial Action Task Force (FATF) high-risk jurisdiction lists are, in function, typed negatives: they define what is excluded.

In software documentation, API deprecation is a temporal negative. A function valid in v2.0 and deprecated in v3.0 embeds identically in both versions. Zhou and Walker (2016) show that deprecation is non-linear; functions are deprecated, un-deprecated, and re-deprecated without predictable sequence. Sawant et al. (2018) confirm that developers do not follow deprecation signals, be it because the signals are buried in changelogs or because the retrieval system does not surface them.

The general principle: any domain where documents can be topically similar but contextually inapplicable needs the properties described in this paper.

6.3 Toward "negative aboutness"

Information science has a rich theory of "aboutness," what a document is about (Hjørland, 2001). There is no corresponding theory of what a document is explicitly *not* about. Current standards provide no formal mechanism for exclusions. A Medical Subject Headings (MeSH) scope note may say "do not confuse with [X]" in natural language, but there is no

structured field encoding this exclusion. ISO 25964 and the Simple Knowledge Organization System (SKOS) offer no `exclusionNote` or equivalent.

The negative evidence taxonomy proposed in this paper (scope, sibling, temporal) is a step toward formalising negative aboutness for retrieval systems. Two traditions of negation work inform but do not address the problem. In clinical NLP, Harkema et al. (2009) extended negation detection with ConText to include temporality, a compound negative signal (negated AND historical), which is structurally analogous to CRANE's intersection of typed negatives and temporal awareness. In information retrieval, Boolean negation operators (NOT, AND NOT) have been available since early systems (Manning et al., 2008; Baeza-Yates and Ribeiro-Neto, 2011), and faceted search allows exclusion filters at query time. The distinction from CRANE is fundamental: IR negation is a query-time operator applied by the user; CRANE's typed negatives are persistent metadata attached at ingestion time, with per-type penalty weights integrated into the merge scorer. The user does not need to know which exclusions apply; the system detects and encodes them. We leave the full formalisation to information scientists but note the gap. That a concept so fundamental to retrieval quality has no formal representation in the field's standard vocabularies is, in our view, worth naming.

6.4 Limitations

The evidence base in Section 3 is drawn from published evaluations across legal, medical, and financial retrieval. Our synthesis of that evidence is only as strong as the underlying studies. Where the originating authors report limitations (sample sizes, domain-specific corpora, evaluation metrics that may not transfer), those limitations propagate to our argument. Independent replication of the failure mode analysis on new corpora would strengthen the foundation.

The negative evidence taxonomy captures the three failure modes most extensively documented in the current literature. It is not claimed to be exhaustive. Other types are plausible *de lege ferenda*: jurisdictional hierarchy negatives (applicable law but wrong court level), conditional negatives (applicable only if a factual predicate holds), quantitative negatives (correct provision but wrong threshold amount). Whether three types suffice or whether domain-specific deployments surface additional types is an empirical question that the taxonomy's typed structure is designed to accommodate; adding a new negative type requires a detection rule, a merge scorer weight, and a generation pattern, not an architectural change.

This paper specifies CRANE's functional architecture (the information flow, the mechanism interfaces, and the properties each mechanism must satisfy) rather than a single implementation. That separation is deliberate: the contribution is the architectural requirement (Properties 1–3 and the feedback loop connecting them), not any particular parameter set. Companion technical documentation addresses detection algorithms, scorer configurations, and deployment specifications. Performance characteristics will vary across implementations; the properties are invariant.

The running example is drawn from tax law. The cross-domain evidence in Section 3 (medical retrieval, financial services, software documentation) and the applications in Section 6.2 suggest the failure modes are structural rather than domain-specific. Formal generalisation to new domains, particularly those where the semantic similarity between correct and incorrect answers is less pronounced, requires empirical validation on domain-specific corpora.

7 Conclusion

Retrieval-augmented generation has made substantial progress in reducing hallucination and improving factual grounding. But the dominant paradigm (rank by positive relevance, retrieve the top-k, generate) has a structural blindspot: it has no way to represent what documents do not answer.

This paper has shown, through evidence from legal, medical, and financial domains, that this blindspot produces the most dangerous errors at the highest confidence levels. We have derived three properties that any retrieval system for high-stakes domains must satisfy: typed negative evidence, calibrated confidence, and confidence-gated generation. We have presented CRANE as one architecture satisfying these properties and proposed a taxonomy of negative evidence types (scope, sibling, temporal) as a first step toward formalising negative aboutness.

The contribution is twofold: the identification of properties that high-stakes retrieval systems must satisfy, and a functional architecture demonstrating their joint satisfiability. The architecture produces three measurable technical effects (reduced false-positive retrieval through typed negative evidence, calibrated relevance probabilities through integrated merge scoring, and differentiated generation through confidence gating) that no existing system achieves in combination. Negative aboutness, what a document is explicitly not about, has no formal representation in current information retrieval standards. That gap is worth closing.

Our claims are falsifiable, and we invite empirical evaluation. Four predictions are stated in Section 6. If a positive-only system matches CRANE on false-positive rates, calibration, generation quality, and temporal accuracy, the argument is wrong.

What doesn't match matters more, because the most dangerous answer is not the one that scores low, but the one that scores high and is wrong.

References

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2024). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the Twelfth International Conference on Learning Representations* (*ICLR 2024*). https://arxiv.org/abs/2310.11511

Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search*. 2nd ed. Addison-Wesley.

Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., and Abdelrazek, M. (2024). Seven Failure Points When Engineering a Retrieval Augmented Generation System. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering — Software Engineering for AI* (*CAIN 2024*), pp. 194–199. https://doi.org/10.1145/3644815.3644945

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '21*), pp. 610–623. https://doi.org/10.1145/3442188.3445922

Berberich, K., Bedathur, S., Neumann, T., and Weikum, G. (2007). A Time Machine for Text Search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 2007*), pp. 519–526. https://doi.org/10.1145/1277741.1277831

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1), 1–3.

Bruch, S., Gai, S., and Ingber, A. (2023). An Analysis of Fusion Functions for Hybrid Retrieval. *ACM Transactions on Information Systems*, 42(1), 1–35. https://doi.org/10.1145/3596512

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, B. G. (2001). A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5), 301–310.

Cohen, D., Mitra, B., Lesota, O., Rekabsaz, N., and Eickhoff, C. (2021). Not All Relevance Scores are Equal: Efficient Uncertainty and Calibration Modeling for Deep Retrieval Models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR 2021*).

Covington, P., Adams, J., and Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (*RecSys '16*), pp. 191–198.

Daivam, E. (2025). Reducing False Positives in Retrieval-Augmented Generation (RAG) Semantic Caching: A Banking Case Study. *InfoQ*.

De Martim, H. (2025). An Ontology-Driven Graph RAG for Legal Norms: A Structural, Temporal, and Deterministic Approach. In *Legal Knowledge and Information Systems*. IOS Press.

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (*IJCAI 2001*), Vol. 2, pp. 973–978.

Feldman, P., Foulds, J. R., and Pan, S. (2024). RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots. *arXiv preprint arXiv:2403.01193*.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), 1189–1232.

Gade, A., Jetcheva, J. G., and Trivedi, H. (2024). It's About Time: Incorporating Temporality in Retrieval Augmented Language Models. *arXiv preprint arXiv:2401.13222*.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning* (*ICML 2017*), PMLR Vol. 70, pp. 1321–1330.

Harkema, H., Dowling, J. N., Thornblade, T., and Chapman, W. W. (2009). ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *Journal of Biomedical Informatics*, 42(5), 839–851.

Hindi, M., Mohammed, L., Maaz, O., and Alwarafy, A. (2025). Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access*, 13.

Hjørland, B. (2001). Towards a Theory of Aboutness, Subject, Topicality, Theme, Domain, Field, Content...and Relevance. *Journal of the American Society for Information Science and Technology*, 52(9), 774–778.

Huwiler, D., Stockinger, K., and Furst, J. (2025). VersionRAG: Version-Aware Retrieval-Augmented Generation for Evolving Documents. *arXiv preprint arXiv:2510.08109*.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. (2023). Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2023*), pp. 7969–7992.

Kalra, R., Wu, Z., Gulley, A., Hilliard, A., Guan, X., Koshiyama, A., and Treleaven, P. C. (2024). HyPA-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications. In *Proceedings of the 1st Workshop on Customizable NLP* (*CustomNLP4U*), *EMNLP 2024*, pp. 237–256.

Kanhabua, N., Blanco, R., and Nørvåg, K. (2015). Temporal Information Retrieval. *Foundations and Trends in Information Retrieval*, 9(2), 91–208.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W. (2020). Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 6769–6781.

Khattab, O. and Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR '20*), pp. 39–48.

Kim, H., Sohn, J., Gilson, A., et al. (2025). Rethinking Retrieval-Augmented Generation for Medicine: A Large-Scale, Systematic Expert Evaluation and Practical Insights. *arXiv preprint arXiv:2511.06738*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems* (*NeurIPS*), Vol. 33, pp. 9459–9474.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV 2017*), pp. 2980–2988.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. (2024). Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *arXiv preprint arXiv:2405.20362*.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McMahan, H. B., Holt, G., Sculley, D., et al. (2013). Ad Click Prediction: A View from the Trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '13*), pp. 1222–1230.

Meghwani, H., Agarwal, A., Pattnayak, P., Patel, H. L., and Panda, S. (2025). Hard Negative Mining for Domain-Specific Retrieval in Enterprise Systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (*ACL 2025*), Industry Track, pp. 1013–1026.

Microsoft Health and Life Sciences. (2025). Harnessing AI's Potential in Healthcare: Overcoming Benchmarking Challenges. *Microsoft Tech Community Blog*.

Newman-Toker, D. E., Nassery, N., Schaffer, A. C., Yu-Moe, C. W., Clemens, G. D., Wang, Z., Zhu, Y., Saber Tehrani, A. S., Fanai, M., Hassoon, A., and Siegal, D. (2023). Burden of Serious Harms from Diagnostic Error in the USA. *BMJ Quality & Safety*, 33(2), 109–120.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. In *Proceedings of the 22nd International Conference on Machine Learning* (*ICML 2005*), pp. 625–632.

Omar, M., Agbareia, R., Glicksberg, B. S., Nadkarni, G. N., and Klang, E. (2025). Benchmarking the Confidence of Large Language Models in Answering Clinical Questions: Cross-Sectional Evaluation Study. *JMIR Medical Informatics*, 13, e66917.

Ozaki, S., Kato, Y., Feng, S., et al. (2024). Understanding the Impact of Confidence in Retrieval Augmented Generation: A Case Study in the Medical Domain. *arXiv preprint arXiv:2412.20309*.

Penha, G. and Hauff, C. (2021). On the Calibration and Uncertainty of Neural Learning to Rank Models for Conversational Search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (*EACL 2021*), pp. 160–170.

Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, MIT Press, pp. 61–74.

Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., and Wang, H. (2021). RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL 2021*), pp. 5835–5847.

Rashkin, H., Nikolaev, V., Lamm, M., et al. (2022). Measuring Attribution in Natural Language Generation Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (*ACL 2022*), Vol. 1, pp. 7054–7066.

Reuter, M., Lingenberg, T., Liepina, R., Lagioia, F., Lippi, M., Sartor, G., Passerini, A., and Sayin, B. (2025). Towards Reliable Retrieval in RAG Systems for Large Legal Datasets. In *Proceedings of the Natural Legal Language Processing Workshop 2025* (*NLLP 2025*), pp. 17–30.

Sawant, A. A., Robbes, R., and Bacchelli, A. (2018). On the Reaction to Deprecation of Clients of 4+1 Popular Java APIs and the JDK. *Empirical Software Engineering*, 23, 2158–2197.

Sciavolino, C., Zhong, Z., Lee, J., and Chen, D. (2021). Simple Entity-Centric Questions Challenge Dense Retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (*EMNLP 2021*), pp. 6138–6148.

Sculley, D., Holt, G., Golovin, D., et al. (2015). Hidden Technical Debt in Machine Learning Systems. In *Advances in Neural Information Processing Systems* (*NeurIPS*), Vol. 28, pp. 2503–2511.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803.

Su, Y., Lu, P., Pan, B., Wen, Y., and Liu, Y. (2024). DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of Large Language Models. In *Proceedings*

of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), pp. 11328–11345.

Wang, S., Zhuang, S., and Zuccon, G. (2021). BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '21)*.

Wang, X., Chi, J., Tai, Z., et al. (2025). FinSage: A Multi-aspect RAG System for Financial Filings Question Answering. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*.

Weller, O., Boratko, M., Naim, I., and Lee, J. (2026). On the Theoretical Limitations of Embedding-Based Retrieval. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR 2026)*.

Yan, L., Qin, Z., Wang, X., Bendersky, M., and Najork, M. (2022). Scale Calibration of Deep Ranking Models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022)*, pp. 4300–4309.

Yan, S.-Q., Gu, J.-C., Zhu, Y., and Ling, Z.-H. (2024). Corrective Retrieval Augmented Generation. *arXiv preprint arXiv:2401.15884*.

Yousuf, R. B., Xu, S., Sharma, M., Neeser, A., Latimer, C., and Ramakrishnan, N. (2026). Utilizing Metadata for Better Retrieval-Augmented Generation. In *Proceedings of the 48th European Conference on Information Retrieval (ECIR 2026)*.

Zhang, R., Xu, Y., Xiao, Y., Zhu, R., Jiang, X., Chu, X., Zhao, J., and Wang, Y. (2025). KnowPO: Knowledge-aware Preference Optimization for Controllable Knowledge Selection in Retrieval-Augmented Language Models. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI 2025)*.

Zhao, X., Liu, S., Yang, S.-Y., and Miao, C. (2025). MedRAG: Enhancing Retrieval-Augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*.

Zheng, L., Guha, N., Arifov, J., Zhang, S., Skreta, M., Manning, C. D., Henderson, P., and Ho, D. E. (2025). A Reasoning-Focused Legal Retrieval Benchmark. In *Proceedings of the 4th ACM Symposium on Computer Science and Law (CSLAW '25)*.

Zhou, J. and Walker, R. J. (2016). API Deprecation: A Retrospective Analysis and Detection Method for Code Examples on the Web. In *Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE 2016)*, pp. 266–277.