Governance-Aware Semantic Interpretation in Legal AI

An Evaluation of Disambiguation-Driven Reliability, Traceability, and Operational Trustworthiness

Executive Summary

Legal AI systems are increasingly being deployed in contexts where the precise meaning of a question — and the meaning the system commits to in its answer — has real downstream consequences. A loan agreement question about "interest" can be read as security interest, beneficial interest, rate of interest, accrued interest, or conflict of interest. Each reading invokes a different body of law, different document provisions, and different operational follow-up. When a legal AI system answers such a question without first committing to a single sense, the result is frequently fluent prose that hedges across multiple legally distinct interpretations. This is not a presentation defect. It is a governance defect.

This whitepaper documents an evaluation framework — the Answer Evaluation Application (AEA) — designed to measure how much reliability and operational trustworthiness a legal AI system gains when it integrates semantic ambiguity detection, explicit sense selection, and provenance-aware retrieval into its reasoning pipeline. The framework scores answers across nine dimensions spanning semantic accuracy, interpretive transparency, traceability, legal craft, and hallucination control. Two answer streams are compared for each question: a **raw** answer generated without disambiguation, and a **disambiguated** answer generated after the system explicitly resolved the legal sense of each ambiguous term.

The evaluation reported here covers 15 legal questions across a representative ambiguity-severity range (high, moderate, low) drawn from a single securities loan agreement. Across all 15 questions, disambiguated answers averaged 9.06 versus 7.16 for raw answers on a 0–10 scale, a 27% lift in mean composite score. The lift is strongly correlated with ambiguity severity: on high-ambiguity questions, the gap widens to roughly 4.1 points; on moderate- and low-ambiguity questions, the gap shrinks to under 1 point. The two largest contributors to the gap are *disambiguation audit trail* (where raw answers leave no record of the interpretive choice) and *traceability and provenance integrity* (where raw answers structurally have no claim-level grounding). Legal craft dimensions — precision of terminology, completeness of operative coverage, practical usability — show meaningful but more modest improvement.

We do not claim that disambiguation "solves" legal AI, that raw systems are unusable, or that interpretation-aware systems are universally superior. The evidence is more specific: **governance-aware semantic interpretation materially improves reliability and operational trustworthiness in ambiguity-sensitive legal workflows**, and the improvement is largest where it is most consequential — questions whose surface wording admits multiple legally distinct readings. For enterprise legal deployment, the operational

implication is that ambiguity resolution is not a quality-of-prose feature; it is a control mechanism.

1. Introduction

Legal language is structurally ambiguous in ways that everyday discourse is not. A single word frequently anchors competing doctrines, each with its own statutory framework, case law, and operational consequences. The word *interest* is the canonical example: in a single contract it can refer to a security interest under Article 9 of the UCC, a beneficial interest under trust law, a rate of interest under lending regulation, accrued interest under accounting practice, or a conflict of interest under fiduciary doctrine. These are not shades of the same concept. They are different legal objects.

A legal practitioner reading a contract resolves these ambiguities continuously and largely without conscious effort, anchored by context, domain expertise, and discipline trained over years. A general-purpose large language model, by contrast, often does not commit to a single reading at all. It hedges. It produces answers that gesture at multiple senses simultaneously, or that pick a sense implicitly without disclosing the choice. The output reads competent — sometimes more comprehensive than a focused human answer — but is operationally unreliable: a reviewer cannot tell which sense the system was actually addressing, cannot reproduce the reasoning that led to the choice, and cannot defend the answer against an alternative reading without doing the interpretive work over from scratch.

This problem has been treated as a quality concern in much of the legal AI literature. We argue that it is more accurately framed as a **governance** concern. A legal AI system that cannot disclose which interpretation it committed to, why it committed to it, and what evidence supports the resulting answer is a system that cannot be audited, reviewed, or relied upon in regulated workflows. Answer correctness, in isolation, is insufficient.

The AEA framework was built to make these properties measurable. It scores legal AI answers not only on whether the conclusion is right, but on whether the interpretive choice is **visible**, **traceable**, **consistent**, and **resistant to drift**. The intent is to give procurement, governance, and compliance leaders a way to compare systems on the attributes that matter for defensibility, not just on the attributes that matter for fluency.

This whitepaper presents the framework, its evaluation methodology, and the results obtained from a 20-question evaluation comparing raw and disambiguated answers against a single source contract. We discuss the patterns the results reveal, the enterprise implications for legal AI procurement and deployment, and the limitations of the current evaluation. We close with a discussion of where the framework should be extended next.

2. The Problem of Semantic Ambiguity in Legal AI

Legal drafting deliberately uses words that, when isolated from context, have multiple legally distinct meanings. This is not a quirk; it is a feature of how doctrine accumulates. A statutory term incorporated into one regime acquires a defined meaning there. The same word, used in a different regime, acquires a different defined meaning. Over time the

language develops a layered polysemy that practitioners navigate by tracking domain context — but that a general-purpose language model has no native mechanism to navigate.

Six terms recur in the ambiguity literature, and we use them throughout this work as representative cases:

- **interest** — security interest (UCC Article 9), beneficial interest (trust/estate law), interest rate (lending), accrued interest (accounting), conflict of interest (fiduciary doctrine).
- **security** — security interest (secured transactions), security as a financial instrument (securities law), physical security (operational), security as collateral (contract).
- **default** — contractual default (event triggering remedies), default as fallback (operational/system default), default as failure (general).
- **margin** — collateral margin (financial), margin as buffer (commercial), formatting margin (document craft).
- **instrument** — legal instrument (a written document with legal effect), financial instrument (a tradable contract), instrument as means (general).
- **party** — contracting party (named in an agreement), political party, party as gathering (general).

The downstream operational consequences of getting the resolution wrong vary in severity. In some cases the wrong reading produces an answer that is merely irrelevant. In others — and this is where legal AI is most exposed — the wrong reading produces an answer that is internally coherent, fluent, plausible-sounding, and substantively wrong in ways that would cause material harm if relied upon.

The risk is amplified by three features of how legal AI is typically deployed:

1. **Reviewer fatigue.** Legal reviewers operating under time pressure tend to spot-check rather than fully re-derive AI outputs. An answer that *reads* competent often passes review even when it is built on a wrong interpretive choice.
2. **Aggregate workflows.** Legal AI outputs frequently feed into downstream artifacts (memos, briefs, due-diligence reports) where the interpretive choice becomes invisible. Once embedded, errors are hard to trace back to their source.
3. **Auditing asymmetry.** Reviewers who *agree* with an AI's chosen interpretation typically do not surface the alternative readings the system implicitly rejected. The interpretive choice, once made, becomes a silent assumption.

A system that does not surface its interpretive commitments, then, does not merely produce occasional bad outputs. It produces a class of outputs that are systematically difficult to govern.

3. Governance-Aware Interpretation Architecture

A governance-aware legal AI system addresses the ambiguity problem upstream of answer generation. Rather than treating interpretation as an emergent property of the language

model's output, it treats interpretation as an explicit, recorded step in the reasoning pipeline. The architecture has five components, each of which produces an auditable artifact:

1. Ambiguity detection. The system identifies which terms in the question are polysemous within the relevant legal domain. This is not a general-language thesaurus lookup. It is keyed to a curated legal ontology in which each polysemous term has a defined set of candidate senses, each with a domain tag and definition.

2. Legal sense classification. For each ambiguous term, the system enumerates the candidate senses, scores each candidate against the document context (and where relevant, the broader question framing), and surfaces the ranking. The ranking is itself an auditable artifact — it records what was considered, not just what was chosen.

3. Interpretation selection. From the candidate ranking, the system either selects a sense automatically (recording a confidence score and the dominance margin against the next candidate) or escalates to a human reviewer when confidence is below threshold. The choice is logged with the reasoning that drove it.

4. Interpretation-aware retrieval. Document retrieval is then keyed to the selected sense, not to the literal question. The retrieval query is rewritten to bake in the chosen interpretation, producing what we call an *augmented query*. The augmented query is itself recorded so the retrieval pathway is reproducible.

5. Provenance-aware answering. The answer is synthesized from retrieved leads with claim-level grounding: each substantive claim is traced to one or more specific source passages (leads), and the claim-level grounding flag is preserved in the trace. The result is an answer that can be audited claim by claim against source artifacts.

Each of these components produces a trace artifact — a structured record of what was considered, what was selected, and what evidence supported the decision. The collected artifacts are what we mean by *traceability*. A legal reviewer presented with the answer, the augmented query, the ambiguity resolution log, and the claim-level source mapping can reconstruct the reasoning pathway end-to-end. A reviewer presented only with the answer cannot.

The architecture is not a prompting technique. It is a control pattern. The substantive contribution is not that the LLM produces better answers; it is that the system produces *defensible* answers — answers accompanied by the evidentiary record needed to support legal review, regulatory audit, and reproducibility.

4. Semantic Infrastructure and Knowledge Graph Governance

The interpretation pipeline described in Section 3 rests on a semantic infrastructure that the preceding sections have so far taken for granted. The architecture refers to a "curated legal ontology" for sense candidates, to document context that scores those candidates, and to provenance-linked claims grounded in source passages — all of which assume an underlying knowledge representation rich enough to support those operations. This section names that infrastructure and explains its role. It is not the headline contribution of

this paper; that contribution is governance-aware interpretation. But governance-aware interpretation is not implementable without semantic infrastructure capable of supporting it, and the choice of that infrastructure has direct consequences for the auditability, explainability, and reproducibility properties measured in later sections.

4.1 Why embeddings alone are insufficient

Modern retrieval-augmented architectures typically rely on vector embeddings — high-dimensional numerical representations of words, phrases, and passages — as their primary semantic substrate. Embeddings are well suited to many tasks: they permit semantic similarity search, generalize across surface forms, and require no hand curation. For governance-aware interpretation in legal AI, however, they are insufficient on their own. Four properties of embedding-only architectures limit their value as the substrate for the operations described in Section 3.

Latent semantic blending. An embedding for the word "interest" is a single vector representing some weighted average of the senses the embedding model has seen. The senses are not separable inside the vector; the model has no internal handle that says "this corresponds to security interest as opposed to interest rate." Retrieval performed against this embedding can match passages from any of the senses, and the system has no internal handle on which sense it is currently retrieving under.

Weak interpretability. Embedding-based decisions cannot be explained at the level of concepts. If a passage was retrieved, the only available explanation is that it had high cosine similarity to the query embedding. That is not an account a legal reviewer can audit. The reviewer cannot ask "what concept was this match made on?" because there is no concept; there is only a vector neighbourhood.

Unstable legal meaning resolution. Identical or near-identical surface queries can produce different retrieved passages across runs, across model versions, and across embedding updates. The instability is not a bug; it is a property of operating against a continuous latent space. For governance-aware workflows this is corrosive: a reproducible answer requires that the system commit to the same interpretation on the same input, which is not a property the embedding layer naturally provides.

Lack of explicit legal-concept anchoring. Legal concepts have names, definitions, and relationships — security interest, beneficial interest, attachment, perfection, priority. These are not noise in the system; they are the objects the law operates on. An embedding architecture has no first-class representation of these objects. It has only representations of the words used to refer to them, mixed together with all other contexts in which those words appear.

4.2 The role of the knowledge graph

A semantic knowledge graph supplies the layer of representation that embedding-only architectures lack. In our system, the graph is a curated structure that models legal concepts and the relationships among them. It serves six related functions that together support the interpretation pipeline.

It models ambiguous legal concepts. Each polysemous term — interest, security, attachment, consideration — is represented not as one entity but as a set of distinct concept nodes, each with a domain tag, a canonical definition, and a stable identifier. Security Interest and Beneficial Interest are different nodes in the graph; the system can address either explicitly.

It maps candidate senses. When a query contains an ambiguous term, the graph supplies the candidate set against which scoring occurs. The candidates are not generated on the fly from the corpus; they are looked up against the graph's ontology, ensuring consistency across queries and across time.

It encodes legal-domain relationships. The graph carries domain-specific edges — granted_by, attaches_to, created_by, requires, limits — that capture the structure of legal reasoning around each concept. These relationships are what permit the coverage-gap analysis described in Section 6: the system can identify concepts adjacent to the selected sense that should appear in the answer based on graph structure, and flag their absence.

It constrains the interpretation space. The graph defines what counts as a legitimate sense for a given term in a given domain. The system cannot resolve "interest" into a sense that is not in the graph; this prevents the model from inventing readings during generation. Constraint is a feature, not a limitation: it is the mechanism that makes interpretive selection auditable.

It supports traceable semantic commitments. Each concept node has a stable identifier that can be recorded in the trace artifact. When the system commits to Security Interest, the trace records the concept ID — not just the string — which means the commitment is recoverable across runs and resolvable against the graph at any later time.

It enables interpretation-aware retrieval orchestration. The graph-anchored sense selection drives the augmented query (Section 6) that constrains retrieval. Retrieval is keyed not to the surface query but to the resolved concept; the retrieved passages are scored not by embedding similarity alone but by their alignment with the structured concept the system has committed to.

4.3 The governance advantage

The knowledge graph in this architecture is not primarily a retrieval-quality optimization. Most discussions of knowledge graphs in retrieval frame them as a relevance lever — a way to improve the precision or recall of retrieved passages. That framing understates what the graph contributes in a governance-sensitive context. The graph's primary value in this system is enabling four properties that retrieval-quality metrics do not directly measure.

Auditability. Every interpretive commitment the system makes can be tagged to a graph entity with a stable identifier. A reviewer can ask, after the fact, which concept did the system commit to here, and receive a determinate answer — not a vector neighbourhood, but a named node with a definition.

Explainability. The graph carries the definitions, the domain tags, and the relationships that let the system articulate why a candidate was selected — Security Interest is the

secured-transactions reading; the document context is a securities loan agreement; this is the dominant candidate by the ontology's structure. Without the graph, the explanation reduces to "the model preferred this passage."

Provenance alignment. The trace artifact links claims to source passages and links source passages to concepts via the graph. A challenge to a specific claim can be resolved at any layer: which passage, which concept, which sense. The provenance chain is not a single layer of text-to-text linking; it is a structured graph traversal that exposes the reasoning at each step.

Semantic reproducibility. Given the same query and the same graph, the system commits to the same concept. The answer text may vary on the surface, but the interpretive pathway — which concept was selected, which were rejected, which document passages were considered relevant under that concept — is reproducible because the graph is reproducible. Reproducibility at this layer is what makes the rest of the governance argument operationally meaningful.

4.4 Relationship to proprietary methods

The graph itself, and the methods by which it drives the interpretation pipeline, embody proprietary semantic governance methods developed for this system. The specifics — how the ontology is curated, how candidates are scored against document context, how the dominance margin and escalation thresholds are tuned, how the augmented query is constructed from the selected concept — are covered by patent-pending interpretation orchestration methods and ontology-backed provenance systems. This whitepaper does not describe those mechanics in detail; the evaluation reported in later sections measures the behavioural properties of the system that result from them. The architectural pattern is generalizable; the specific instantiation is proprietary.

4.5 The strategic claim

It is worth being explicit about what is and is not claimed by the framework. Using a knowledge graph in retrieval is not novel; knowledge graphs have been used in information retrieval, question answering, and recommendation systems for decades. What is being claimed here is more specific: the use of legal-domain semantic graphs to operationalize governance-aware interpretive reasoning and provenance-aware answer generation. The contribution lies in the binding between the graph, the interpretation pipeline, the audit trail, and the answer — the property that every commitment the system makes is anchored to a graph entity, recorded in the trace, and reproducible across runs. A general-purpose knowledge graph plugged into a conventional retrieval system would not produce this property; a legal knowledge graph used only for query expansion would not produce it either. The property emerges from the integration described in Section 3, made implementable by the infrastructure described here.

5. Why Governance-Aware Interpretation Is Not Traditional RAG

The architecture described above bears a surface resemblance to retrieval-augmented generation (RAG): a question arrives, supporting documents are retrieved, and an answer

is synthesized from the retrieved evidence. A skeptical technical reader could reasonably ask whether this is anything more than RAG with extra metadata. The distinction matters, because it shapes what kind of failure the system is designed to prevent and what kind of evidence it is designed to produce.

Traditional RAG retrieves text. Governance-aware interpretation retrieves interpretively constrained text. The difference is not in the retrieval mechanism itself but in what gets retrieved, what is recorded around the retrieval, and what is exposed to downstream review. Specifically, the architecture adds five capabilities that ordinary RAG does not provide:

1. Semantic governance. Ordinary RAG does not model legal sense selection. When a query contains a polysemous term — *security*, *consideration*, *party*, *instrument* — a conventional retrieval pipeline treats the term as a string to match (lexically or via vector similarity) and lets the language model resolve the ambiguity implicitly during generation, if it resolves it at all. Governance-aware interpretation makes the sense choice explicit: it enumerates the candidates against a curated legal ontology, scores them against the document context, and binds the chosen sense to the rest of the pipeline. The retrieval query is not the user's query; it is the user's query under a specific, recorded interpretation.

2. Interpretation-state management. Ordinary RAG does not preserve interpretive state across the pipeline. Each stage — retrieval, ranking, generation — operates over the surface form of the query and whatever passages it pulled. There is no persistent object that says *we are answering this question under sense X, with candidates Y and Z considered and rejected*. Governance-aware interpretation maintains that object as a first-class artifact: it is the spine of the trace, and every downstream step — retrieval, grounding, answer synthesis — is keyed to it.

3. Ontology-guided retrieval transformation. Ordinary RAG does not transform the retrieval target around an interpretive commitment. Hybrid retrieval and query expansion exist in advanced RAG implementations — synonyms, paraphrases, hypothetical-document embeddings — but they expand the query in the direction of *anything that might be relevant*. Governance-aware interpretation does the opposite: it constrains the retrieval target to the documents and passages consistent with the selected sense. The augmented query is not broader than the user's query; it is narrower and more specific, scoped to a single defended reading.

4. Audit-oriented reasoning architecture. Ordinary RAG does not maintain audit traces in the sense required for legal or regulatory review. Most RAG systems log retrieved chunks and the final answer; some log retrieval scores. Few log what was considered and rejected, the reasoning that drove the selection, the confidence and dominance margins, or the human-in-the-loop escalation pathway. Governance-aware interpretation is built around the audit artifact, not around the answer. The answer is a derivative of the trace; the trace is the primary output.

5. Provenance-linked interpretive commitments. Ordinary RAG does not produce reproducible semantic commitments. Re-running the same query against the same corpus typically yields different retrieved passages and different generated text, because retrieval

is non-deterministic at scale and generation is stochastic. Reproducibility in governance-aware interpretation operates one level up: given the same input, the system commits to the same interpretation, retrieves under the same augmented query, and binds claims to the same source passages. The surface text of the answer may still vary, but the interpretive pathway — what the system understood the question to mean and what it grounded the answer in — is stable and reproducible.

Taken together, these five capabilities describe a different control pattern from RAG, not an extension of it. RAG is an augmentation layer over generation: it tries to make the model's answer better by giving it more relevant context. Governance-aware interpretation is a constraint layer over interpretation: it tries to make the system's reasoning defensible by recording what was understood, what was rejected, and what evidence underwrites each claim. The two are not mutually exclusive — a governance-aware system uses retrieval, and that retrieval can be implemented with RAG techniques — but the locus of value is different. RAG asks *what should the model see?* Governance-aware interpretation asks *what does the model claim to mean, and can a reviewer audit that claim?*

The distinction is strategically important because the failure modes diverge. RAG can fail silently when retrieved evidence is plausibly relevant but interpretively wrong — the user receives a fluent, well-grounded answer to the wrong question. Governance-aware interpretation is designed to surface that failure: if the wrong sense is selected, the selection itself is visible in the trace, and a reviewer can challenge it. The system does not promise to never err; it promises that errors are inspectable. That is the architectural commitment that separates governance-aware interpretation from retrieval augmentation, regardless of how sophisticated the retrieval layer becomes.

6. A Worked Example: "What interest does the lender have?"

This is exactly the kind of question a lawyer would ask without giving any further thought to the meaning of "interest." It is short, fluent, and immediately ambiguous — and a non-disambiguated system will typically answer it without ever surfacing that ambiguity. The trace recorded by the governance-aware system on this exact question, drawn from the live evaluation, illustrates the architecture end-to-end. Every value shown below is verbatim from the system's recorded trace artifact for this query.

Stage 1 — The question. The user-supplied question is recorded as-is:

> what interest does the lender have?

Stage 2 — Ambiguity detection. The system flags "interest" as polysemous within the legal domain. Five candidate senses are enumerated against the curated ontology and scored for contextual relevance through a multi-pass evaluation. The full candidate ranking is recorded:

#	Sense	Domain	Confidence
1	Security Interest	secured_transactions	88%
2	Beneficial Interest	trust_law	79%

3	Interest Rate	lending	76%
4	Accrued Interest	accounting	39%
5	Conflict of Interest	governance	0%

Stage 3 — Sense classification. Each candidate is scored against the document context, producing the relevance ranking shown above. The ranking is itself recorded in the trace — the system can later show not only which sense it selected but how each alternative compared.

Stage 4 — Interpretation selection. The top sense — Security Interest — was the leading candidate. The system did not auto-accept, however: the dominance margin between the next two candidates (Beneficial Interest and Interest Rate) was below the system's escalation threshold, indicating that the alternatives that might be the right reading were too close to distinguish without human confirmation. The flagged reason recorded in the trace is verbatim:

> "Dominance margin 0.809 is below threshold 1.50 ('Beneficial Interest' vs 'Interest Rate' — too close to call)."

Because the alternatives were competitive, the choice was surfaced for human confirmation rather than taken silently. The user confirmed Security Interest. This is the governance check working as designed: when the top sense dominates clearly, the system proceeds automatically; when alternatives are competitive, the choice is escalated. The recorded artifact captures not only the selection but the rejected alternatives (Beneficial Interest, Interest Rate, Accrued Interest, Conflict of Interest), each with its score, so a reviewer can challenge the choice against the alternatives that were considered.

Stage 5 — Augmented query. Under the selected sense, the user's original query is rewritten:

> Original: "what interest does the lender have?"
>
> Augmented: "what security interest does the lender have?"

The augmented query — not the user's original — is what drives retrieval. The system also retains four secondary queries scoped to the rejected senses (one each for Beneficial Interest, Interest Rate, Accrued Interest, and Conflict of Interest), so a reviewer who wants to test an alternative reading can rerun retrieval under it without rebuilding the query from scratch. The secondary queries are not used for the primary answer.

Stage 6 — Retrieved leads. Retrieval under the augmented query returned seventeen leads, all anchored to Section 8 of the agreement and its sub-paragraphs. Two of these were the load-bearing passages for the answer:

> **lead_30** (Section 8, ¶3.3): "The Collateral Delivered by Borrower to Lender, as adjusted pursuant to Section 8 below, shall be security for all Borrower's present and future obligations arising from this Agreement, and Borrower hereby grants Lender a continuing security interest in and a lien upon, and pledges with and assigns to Lender the Collateral, which shall attach upon the Delivery of the

> Collateral to Lender and which shall cease upon the Redelivery of the Collateral to Borrower."
>
> **lead_31** (Section 8): "In addition to the rights and remedies given to Lender hereunder, Lender shall have all the rights and remedies of a secured party under the Personal Property Security Act (Ontario), as amended from time to time."

Stage 7 — Grounded answer. The synthesizer composes the answer from the retrieved leads, prepending an explicit interpretation note and tagging each substantive claim with its supporting lead identifier:

> *Interpretation note: 'interest' was interpreted as 'Security Interest'; alternatives set aside: Beneficial Interest, Interest Rate, Accrued Interest.*
>
> The lender has a continuing security interest and lien over the collateral delivered by the borrower. Specifically: the collateral "shall be security for all Borrower's present and future obligations arising from this Agreement," and the borrower grants the lender "a continuing security interest in and a lien upon, and pledges with and assigns to Lender the Collateral, which shall attach upon the Delivery of the Collateral to Lender and which shall cease upon the Redelivery of the Collateral to Borrower." [lead_30] In addition to contractual rights, the lender has "all the rights and remedies of a secured party under the Personal Property Security Act (Ontario), as amended from time to time." [lead_31]

Stage 8 — What the raw answer looked like. For the same question against the same document, the non-disambiguated pipeline produced a five-bullet answer covering economic interest in collateral and proceeds, indemnification rights, entitlement to distributions on loaned securities, economic interest in distributions on cash collateral, and voting rights as "legal and beneficial owner." The raw answer ended with the following hedge:

> "If you meant a different type of 'interest' (e.g., security interest vs. ownership interest), the document passages supplied do not clearly describe the legal characterization beyond stating that the lender is treated as legal and beneficial owner of the loaned securities for rights and benefits such as voting and distributions."

The two answers are not factually inconsistent — the raw system found real passages and reported real provisions — but only one of them commits to an interpretation, surfaces the alternatives that were considered, and tags each substantive claim to a specific source passage. A reviewer presented with the disambiguated answer can verify the interpretive commitment, audit each claim against the named lead, and trace the reasoning pathway end-to-end. A reviewer presented with the raw answer has to guess which question was actually being answered — and is then asked, by the hedge in the closing paragraph, to do the disambiguation themselves.

What is and isn't recorded in the trace. The single trace artifact captured for this question contains: the original query, the resolved interpretation with confidence and dominance scores, all five candidate senses with their confidence scores, the flagged-

reason explanation, the augmented query, four secondary queries (one per rejected sense), the seventeen retrieved leads with their source-section anchors, a per-conclusion traceability log mapping each output sentence to its contributing lead identifiers, and a coverage-gap report listing concepts adjacent to "security interest" that were expected to appear but did not (notably "debtor" and "loan agreement") so a reviewer can challenge omissions, not only commissions.

This is the artifact that the rest of the paper measures. The dimensional scores reported in Section 11 are evaluator judgments about the properties of this trace and traces like it: whether the sense commitment is consistent, whether the audit trail is complete, whether each claim is grounded, whether the rejected interpretations are visible. The architectural argument made in Sections 3 and 4 reduces to whether artifacts of this shape can be produced reliably, at scale, and with the cost properties enterprise deployment requires. The empirical sections that follow address that question.

7. Interpretive Hedging

The worked example in Section 6 closed with a hedge: "If you meant a different type of 'interest'..." That single sentence, and the multi-sense answer that produced it, illustrates a systematic failure mode of non-disambiguated legal AI systems — one that conventional output-quality assessment is structurally blind to. We name it formally.

7.1 Definition

> ***Interpretive hedging.*** *The tendency of non-disambiguated legal AI systems to produce answers that implicitly blend multiple legally distinct senses to maximize apparent completeness while reducing semantic commitment.*

Three terms in this definition do real work and deserve unpacking. *Implicitly*: the blending is not announced. A system that produces a five-bullet answer covering five possible meanings of "interest" does not say *I am covering five senses because I cannot determine which one you asked about*. It emits all five and lets the reader sort them. *Blends*: the senses are not held distinct in the answer's structure. A reviewer reading the bullets cannot draw a clean line between "this bullet is about security interest" and "this bullet is about beneficial interest"; the senses interweave at the level of clauses and qualifying phrases. *Apparent completeness*: the answer reads as thorough, which is itself the failure mode. Apparent completeness is the property that fools surface-fluency assessment, and it is precisely what makes interpretive hedging difficult to catch in conventional output review.

7.2 Why interpretive hedging arises

The mechanism is not adversarial; it is the default behaviour of a generative model in the presence of unresolved ambiguity. Three forces converge to produce it.

First, the model has no externally enforced commitment to a single sense. Disambiguation, if it happens at all, happens implicitly inside the generation process and is not a recorded step. The model is free to drift across senses sentence by sentence because nothing in the pipeline asks it to declare what "interest" means before it starts writing.

Second, the model has been trained on text in which covering multiple readings reads as helpful. Across general-purpose corpora, hedging on word meaning is innocuous and often positively valuable, as in encyclopedic writing or explanatory prose. The training signal does not distinguish between contexts where hedging serves the reader and contexts where it abdicates the reader's question.

Third, the model is rewarded by downstream signals — user feedback, reviewer reactions, automated quality metrics — that respond to fluency and apparent thoroughness more reliably than to interpretive precision. A five-bullet answer that covers the territory wins on most assessment instruments. A focused answer that picks a sense and declares the alternatives it set aside has fewer surface features to grade well on.

7.3 Why this is a legal-AI-specific problem

In general writing, hedging on word meaning is mostly harmless and sometimes useful. In legal writing, the operative term *is* the thing being decided. Hedging on what "security interest" means is hedging on the legal question itself. The reader is not free to mentally average across senses; they have to choose one for any downstream legal action, and the system has just declined to help them choose.

This is the difference between a legal answer and a legal-looking answer. A legal answer commits to a reading and supports it against the alternatives. A legal-looking answer surveys the senses, sounds informed, and leaves the interpretive labour with the reader — who is now doing the work the AI was supposed to do, under the impression that the AI has already done it.

7.4 Distinguishing interpretive hedging from legitimate hedging

Not all hedging is interpretive hedging, and the distinction matters for both evaluation and design. Three categories should be kept separate.

Epistemic hedging — "I don't know," "the document does not specify," "the record is silent on this point" — is explicit, named, and valuable. A system that flags a genuine information gap performs better than one that fabricates content to fill it. Legitimate epistemic hedging is a feature, not a failure mode.

Legal-interpretation hedging — "this provision is ambiguous and admits two readings, X and Y" — is also explicit and valuable. A system that surfaces genuine textual ambiguity, names the readings, and walks through the legal implications of each is doing the lawyer's work, not deflecting it. The key property is that the alternatives are named and the ambiguity itself is the subject of the answer.

Interpretive hedging is the inverse of both. It is implicit (the hedging is not announced); it is term-level (the ambiguity is in the question's vocabulary, not in the document's text); and the alternatives are blended into the answer rather than named as alternatives. The reader is not told that hedging has occurred. The reader is, in fact, given the impression that the question was understood and answered.

7.5 How interpretive hedging shows up in evaluation

The rubric used in this paper measures interpretive hedging directly through three dimensions: Sense Commitment, Semantic Resolution Accuracy, and Audit Trail Quality. Sense Commitment penalises answers that drift across senses or fail to declare which sense is addressed; Semantic Resolution Accuracy penalises wrong or unstable interpretations; Audit Trail Quality penalises the absence of a recorded selection. Together, these three dimensions produce the large severity-correlated gaps reported in Section 11. Conventional output-quality assessment misses interpretive hedging by design: fluency, grammaticality, topical-coverage, and most groundedness scores treat a multi-sense answer as well-formed. The hedging is invisible to any evaluator that does not specifically ask which sense was committed to, and was the commitment recorded? Governance-aware evaluation requires governance-aware rubrics: the question being measured has to be the right question.

7.6 Implications

The architectural lever that eliminates interpretive hedging is the one this paper has argued for throughout: an external, recorded sense-commitment step in the pipeline. Once the system is required to declare its interpretation before generating its answer, hedging stops being possible — the senses set aside have been named in the trace as alternatives considered and rejected. The hedge, if it would have occurred, becomes an explicit ambiguity finding instead. The procurement and risk consequences are taken up in Sections 7, 12, and 13.

8. False Completeness (the Coverage Illusion)

Section 7 named interpretive hedging as the generation-side failure mode of non-disambiguated legal AI. This section names the evaluation-side counterpart: the systematic mistake by which an interpretively hedged answer is perceived as a more complete, more thorough, more helpful answer than a focused, sense-committed one. We call this *false completeness*, or equivalently, the *coverage illusion*. It is the property of multi-sense answers that makes them look better than they are, and it is the property of conventional assessment that makes them score better than they should. The two phenomena reinforce each other — hedging produces the surface that the illusion rewards; the illusion sustains the procurement preferences that select for hedging systems — and neither can be corrected without addressing the other.

8.1 Definition

> ***False completeness (the coverage illusion).*** *The perception, by reviewers and evaluation instruments, that an answer covering multiple legally distinct senses is broader, more comprehensive, and more helpful than a single-sense answer that commits to one interpretation — when in fact the multi-sense answer reduces operational reliability and increases downstream review burden.*

Three properties of false completeness matter and are worth naming separately. *Surface advantage*: the multi-sense answer looks better on its face — more bullets, more covered

ground, more apparent thoroughness. *Operational disadvantage*: the multi-sense answer is harder to act on, harder to audit, and harder to defend. *Assessment misalignment*: most evaluation instruments reward the surface property and ignore the operational one, so the answer that scores higher is the answer that performs worse.

8.2 Why reviewers and benchmarks reward breadth

Three forces converge to produce the coverage illusion. None of them is malicious; each is the default behaviour of a reasonable evaluation process applied to AI output without governance-aware framing.

First, human reviewers reading a legal AI output under time pressure tend to credit visible coverage. An answer that mentions five aspects of "interest" — economic interest in collateral, indemnification rights, distribution entitlements, voting rights, security interest — feels more thorough than an answer that picks one aspect and goes deep. The cognitive heuristic is sensible in most domains: more covered ground correlates with more knowledge. In domains where the question is asking *which one of these things*, the heuristic inverts — but the reviewer has no reflex that distinguishes the two situations and may not know that a distinction needs to be made.

Second, automated assessment instruments — fluency scores, topical-coverage scores, comprehensiveness metrics, most retrieval-quality benchmarks — were designed for tasks in which broad coverage is genuinely better: summarisation, open-domain question answering, conversational helpfulness. Applied to legal question-answering, those metrics systematically over-reward multi-sense answers. A coverage metric that rewards mentioning more relevant concepts has no built-in way to penalise an answer that mentions five legally distinct concepts when the question asked about one of them.

Third, procurement evaluations of legal AI typically rely on output-sample review by domain experts. The experts are skilled at spotting factual errors and questionable legal reasoning, but they are not trained to detect the absence of an interpretive commitment, because that absence is invisible by definition. A sample-review process that does not include rubric dimensions for sense commitment, audit trail, and traceability will, on average, prefer the system that produces interpretively hedged answers — because those answers will pass the surface-quality check, while the committed answers will read as comparatively thin.

8.3 The breadth-reliability inversion

The core operational fact about false completeness is that breadth and reliability move in opposite directions for interpretively ambiguous legal questions. An answer that covers five senses of "interest" is broader than an answer that commits to one, but it is less reliable in three specific ways.

It is less actionable. A reviewer or a downstream consumer cannot act on five senses; they must pick one. The interpretive labour has been shifted from the system to the user. The user, lacking the ontology and the document context the system had, is generally less well-

positioned to pick than the system was. The answer is broader on its face and narrower in operational value.

It is less auditable. The senses are not held distinct in the answer's internal structure, so a reviewer challenging a specific bullet cannot trace it back to the sense under which it was generated. The reviewer cannot tell whether a particular bullet was a considered claim under one interpretation or a casual mention under another. A challenge to the answer therefore has to be a challenge to the whole answer rather than to a specific claim under a specific reading.

It is less defensible. When the system's output is challenged — in litigation, in regulatory examination, in professional-conduct review — the defending party cannot reconstruct which question the system was actually answering. A multi-sense answer that hedges across security interest, beneficial interest, and interest rate provides no defensible record of the interpretive commitment, because no commitment was recorded. The defence is reduced to *we reviewed the output and it looked reasonable*, which is an exposed position whenever specificity is the question.

8.4 Redesigning assessment to neutralize the illusion

The illusion is not eliminated by exhortation. It is eliminated by changes to the assessment apparatus itself, in three places.

Rubric design. At the per-output level, the rubric should distinguish between "this answer is thorough" and "this answer is correctly focused." Thoroughness is appropriate when the question is broad; focus is appropriate when the question is interpretive. A rubric that does not separate the two will reward the wrong property in the second case.

Worked-example testing. Procurement teams should include interpretively ambiguous queries in their evaluation — questions of the form "what interest does the lender have?" (see Section 6) — and grade systems on the trace artifact produced, not only on the answer text. A confident multi-bullet answer to such a query, with no exposed consideration of which sense was meant, has failed the procurement test regardless of whether the bullets are accurate.

Reviewer training. Training material for legal AI reviewers should include explicit content on the coverage illusion and interpretive hedging. A trained reviewer can apply a second-pass check — did the system tell me which sense it committed to, and did it tell me what it set aside? — that the unaided cognitive heuristic toward thoroughness will not catch.

8.5 The pairing with interpretive hedging

False completeness and interpretive hedging are two sides of the same coin: hedging produces the surface; the illusion rewards it. Neither end is corrected without the other. Architectural sense commitment on the generation side and governance-aware rubric design on the evaluation side together realign breadth and reliability, so that the better answer also scores as the better answer.

9. Evaluation Methodology

The AEA framework scores legal AI answers across nine dimensions. Each dimension has a self-contained scoring prompt that an evaluator model (an LLM operating in a judging role) applies independently for each (question, answer type, dimension) cell. Scores are 0–10 with calibration anchors at each integer band.

A note on intended use. This framework is built for *comparative operational evaluation* of legal AI systems — ranking candidate systems against each other on a rubric chosen by the deploying organization, scored consistently across the candidates — not for *academic benchmark publication*. The distinction matters and shapes how the methodology should be read. A comparative operational evaluation asks whether one system is more reliable than another for a specific deployment context; the score values are meaningful in relation to each other but are not portable across rubrics. An academic benchmark would require properties this framework deliberately does not provide: a fixed canonical dataset, an inter-rater-reliability study quantifying evaluator agreement, calibration against a panel of human experts, a published reproducibility protocol, and a peer-reviewed defence of the rubric's construction. The methodology that follows should be evaluated against the use case it is built to serve, not against benchmark-publication standards it was not designed to meet.

The nine dimensions are:

1. **Semantic Resolution Accuracy** — whether the selected legal sense matches what an experienced reader would conclude from the document context, and whether the answer maintains that sense throughout.
2. **Sense Commitment** — whether the answer commits cleanly to one coherent legal sense without hedging across alternatives.
3. **Disambiguation Audit Trail** — whether a reader can reconstruct why this sense was chosen, through governance metadata, answer prose, or both.
4. **Interpretive Transparency** — whether the interpretive stance is discoverable, treating governance metadata as a transparency artifact.
5. **Traceability & Provenance Integrity** — whether substantive claims are grounded in identifiable source artifacts (document passages, structured leads) and whether citations actually support the attached claims.
6. **Legal Precision** — whether the answer uses legally precise terminology, distinguishes related concepts, and avoids colloquial substitutes for terms of art.
7. **Grounded Completeness** — whether the answer covers the operative provisions directly responsive to the question, scoped to the resolved interpretation, without penalizing focused omission of adjacent material.
8. **Practical Legal Usefulness** — whether a working lawyer reviewing the answer under time pressure would find it immediately actionable.
9. **Overreach & Hallucination Control** — whether the answer stays within what document and trace evidence support, or extends beyond available support.

For each question, the system produces both a **raw** answer (generated directly from the original question without semantic interpretation) and a **disambiguated** answer (generated after ambiguity resolution, with a trace-aware retrieval pathway). Both answers are scored independently against the same rubric. The evaluator model receives the document, the question (annotated with the chosen interpretation for disambiguated cells), the answer text, the rubric scoring prompt, and — for disambiguated answers in trace-aware mode — a structured disambiguation governance block and coverage gap analysis. Raw answers are always scored under blind conditions: the evaluator does not see the trace, so the raw answer is judged on its own merits without the benefit of context the answer-generating system did not have.

Per-dimension scores are weighted and aggregated to produce per-question composite scores for each answer stream, and the resulting score matrix is presented for review. The framework also supports an optional **trust adjustment rule** — a post-evaluation transformation that down-adjusts dimensions whose reliability depends on others (for example, reducing the Practical Legal Usefulness score in proportion to how poorly the answer scored on Traceability & Provenance Integrity). The trust adjustment is view-only; original scores are preserved.

Two predictable objections. A sophisticated reader will raise two objections to this methodology — evaluator subjectivity and rubric circularity. Each is a known cost of the framework's design choices, each is accepted because of the intended use case, and each warrants an explicit response.

Evaluator subjectivity. Using an LLM evaluator rather than a panel of human reviewers introduces measurement noise, since the evaluator is itself a probabilistic system. We accept this cost because consistent scoring at the scale and cost required for repeated operational evaluation is not achievable with human panels in most enterprise procurement processes. The evaluator is treated as a proxy for an experienced legal reviewer applying the calibration anchors; the assumption is flagged for empirical validation in Section 15.

Rubric circularity. The rubric was authored by people who believe semantic governance matters in legal AI, which means it measures what its authors think matters. We accept this cost because the framework's value is in consistency across compared systems on the rubric chosen by the deploying organization, not in external validation of the rubric's construction. A different rubric author would calibrate differently. An organization that disagrees with the rubric's priorities can construct one that reflects its own and apply the same comparative methodology to it; the framework is rubric-agnostic by design.

None of these costs invalidates the comparisons this framework supports. Each would need to be addressed if the framework were positioned for academic publication; none needs to be addressed for the framework to function as a comparative operational tool. The findings below should be read against that use case.

10. Dataset Description

The evaluation reported here covers 15 legal questions, each evaluated against a single source document — a securities loan agreement (with set-off provisions) approximately 38,000 characters in length, drawn from a representative commercial transaction. The questions were stratified by ambiguity severity:

- **High ambiguity (5 questions)** — questions containing at least one term with multiple legally distinct senses that the document context does not unambiguously resolve. Example: *"What interest does the lender have?"* (interest → security / beneficial / rate / accrued / conflict).
- **Moderate ambiguity (5 questions)** — questions where ambiguity exists but the document context substantially narrows the candidate set. Example: *"What attachment rights does the agreement grant?"* (attachment → secured-transactions lien attachment / general document attachment).
- **Low ambiguity (5 questions)** — questions whose intended sense is essentially explicit in the wording or context. Example: *"Are there any liens recorded against the collateral?"*

All 15 questions are real captures from the evaluation framework operating against the source contract. Each was scored across the nine rubric dimensions for both a raw answer and a disambiguated answer.

Table 1 summarises the full dataset.

#	Question (truncated)	Ambiguity	Resolved sense	RAW overall	DIS overall	Δ
1	what interest does the lender have?	High	Security Interest	5.4	9.1	+3.7
2	Has any material default occurred?	Low	Material Adverse Change	8.1	8.6	+0.5
3	Can either party exercise set-off rights immediately after default?	Low	Security Right	7.7	8.2	+0.5
4	Is the margin sufficient to protect the lender's interest?	Moderate	Margin Percentage	8.3	9.3	+1.0
5	Was valid consideration provided for the transfer?	Low	Contractual Consideration	8.0	9.0	+1.0
6	Does the borrower have rights in the collateral after delivery?	Low	Security Right	8.0	9.3	+1.3
7	Is the instrument enforceable after default?	Moderate	Legal Instrument	8.7	9.4	+0.7
8	What notice is required	High	Legal Notice	2.8	8.4	+5.6

	before exercising remedies?					
9	Does the lender retain a security after redelivery?	High	Security Interest	6.6	9.6	+3.0
10	What instrument grants the borrower its security interest?	Moderate	Security Interest	8.6	9.1	+0.5
11	Does the agreement create a first interest in the collateral?	Low	Security Interest	8.8	9.4	+0.6
12	What material representations survive closing?	Moderate	Material Adverse Change	7.5	8.7	+1.2
13	Can the borrower substitute securities without further consideration?	High	Contractual Consideration	4.7	9.3	+4.6
14	What interest survives default?	High	Default	5.5	9.1	+3.6
15	What rights survive termination of the loan?	Moderate	Contractual Right	8.8	9.3	+0.5

Aggregate composites: RAW mean 7.16, DIS mean 9.06, mean Δ +1.90 across all 15 questions.

The dataset is acknowledged to be small and drawn from a single document. We discuss the implications of these constraints in Section 16.

11. Findings & Analysis

11.1 Overall pattern: disambiguation lift is real and severity-correlated

The headline finding is the alignment between ambiguity severity and the size of the disambiguation lift. Table 2 shows mean RAW and DIS scores stratified by severity.

Ambiguity severity	N	RAW mean	DIS mean	Mean Δ
High	5	4.99	9.09	+4.10
Moderate	5	8.38	9.18	+0.80
Low	5	8.11	8.90	+0.79
All	15	7.16	9.06	+1.90

The high-ambiguity tier shows almost a five-point gap on a ten-point scale — a substantial reliability improvement on the questions that, by design, are most exposed to interpretive failure. The low-ambiguity tier shows under a one-point gap; on questions where the intended sense is essentially explicit, the raw system performs nearly as well as the

disambiguated system. This is the expected pattern if disambiguation is doing what it claims to do: helping where help is needed, contributing marginally where it is not.

We treat the severity-correlation as the strongest evidence in the evaluation. A flat lift across all severity tiers would be consistent with a generic prompt-engineering benefit (the disambiguation pipeline producing better-written answers for orthogonal reasons). The observed pattern — large lift where ambiguity is high, small lift where ambiguity is low — is consistent with the disambiguation mechanism doing the work the architecture claims it does.

11.2 Aggregate per-dimension radar (all 15 questions)

Figure 1 plots the mean RAW and mean DIS scores across all 15 questions on a per-dimension radar. Each spoke is one rubric dimension; each polygon is one answer stream. The disambiguated polygon sits at or near the outer ring on every dimension. The raw polygon collapses sharply inward on Traceability & Provenance, and is meaningfully indented on the four disambiguation-sensitive dimensions (Semantic Resolution Accuracy, Sense Commitment, Disambiguation Audit Trail, Interpretive Transparency). On the three legal-craft dimensions (Legal Precision, Grounded Completeness, Practical Legal Usefulness), the two polygons are closer to parallel — both streams produce competent legal writing.

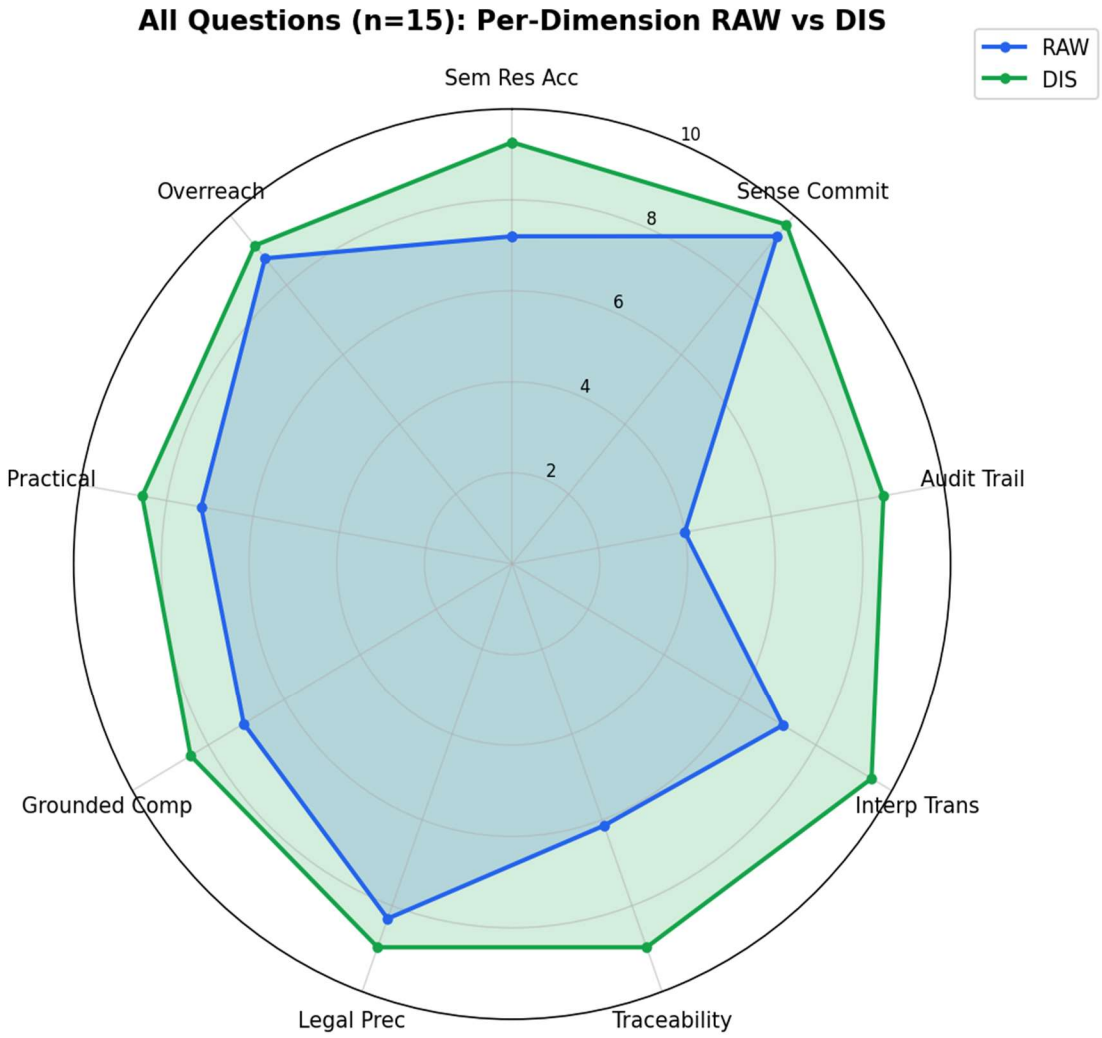

Figure 1. All 15 questions: RAW vs DIS per-dimension means. The disambiguated stream traces the outer ring; the raw stream collapses inward on the governance-sensitive spokes.

The visual is the entire procurement story compressed into one figure. A buyer comparing two systems on output fluency would see them as roughly substitutable — both polygons reach the upper bands on Legal Precision and Practical Legal Usefulness. A buyer comparing them on the full governance shape sees one of them as a reliable substrate for regulated workflows and the other as one that cannot support claim-level audit.

11.3 Severity-stratified radars

The aggregate shape masks an important second-order finding: the size of the disambiguation gap is not uniform across questions. It is strongly correlated with ambiguity severity. The next three figures plot the same nine dimensions but filtered to each severity tier in turn.

High-ambiguity tier (n=5). The polygons here are the most dramatic of the four views. The raw polygon collapses to its smallest footprint across the governance dimensions —

Disambiguation Audit Trail (2.4), Traceability & Provenance (3.2), Semantic Resolution Accuracy (3.4), Grounded Completeness (3.8), and Practical Legal Usefulness (4.8) all sit at or below 5 on a 10-point scale. Sense Commitment is higher at 9.0 across this tier — raw answers still commit to a sense even when the sense is wrong, which makes the commitment dimension a poor discriminator on this tier and reinforces the architectural argument that the audit-trail dimensions are the load-bearing ones. The disambiguated polygon, by contrast, holds near the outer ring on all dimensions. This is the tier where the disambiguation architecture is doing the most work. It is also the tier where a non-disambiguated system is operationally least safe.

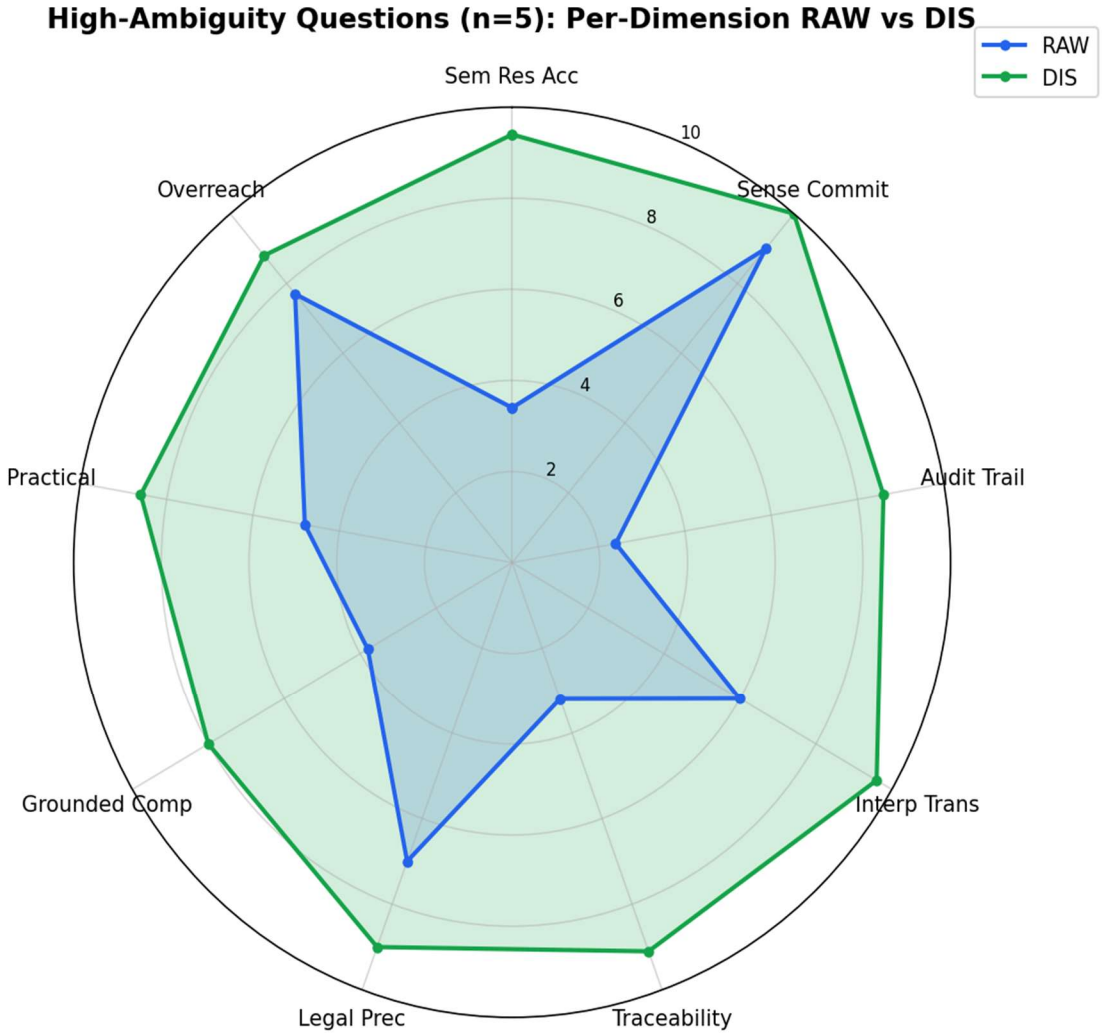

Figure 2. High-ambiguity questions (n=5): the disambiguation lift is largest where the question's surface wording admits multiple legally distinct readings.

Moderate-ambiguity tier (n=5). The raw polygon recovers substantially relative to the high-ambiguity view: most of the raw polygon sits in the 7–10 band rather than the 2–4 band, with Disambiguation Audit Trail (4.6) as the polygon's only material indentation. Traceability has lifted to 8.4 on this tier (vs 3.2 on the high-ambiguity tier), narrowing the architectural gap considerably. The DIS polygon still leads on every spoke, but the

magnitude of the lift is now small enough that the two polygons run closely parallel. The middle tier is where the trade-off conversation gets most interesting: a working reviewer might rate raw answers as serviceable here, while still being unable to audit them against source artifacts.

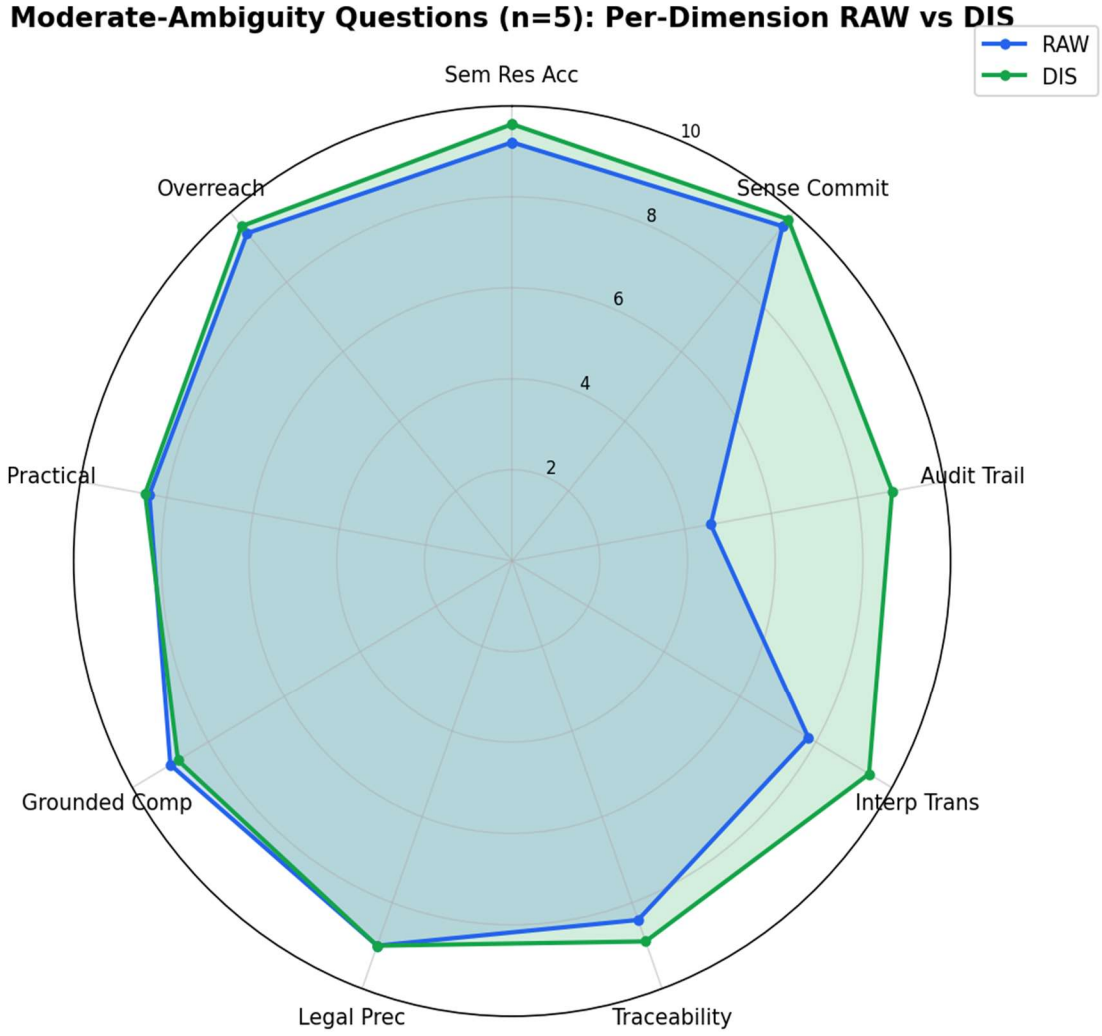

Figure 3. Moderate-ambiguity questions (n=5): raw answers recover on most dimensions but the Disambiguation Audit Trail and Traceability gaps persist.

Low-ambiguity tier (n=5). This is where the conventional procurement question — "do these answers look good?" — collapses to "they look the same." The two polygons run closely parallel across most dimensions; the one remaining material gap is Disambiguation Audit Trail (raw 5.0 vs DIS 8.4), an architectural artifact rather than a writing failure. Traceability is essentially closed on this tier (raw 6.8 vs DIS 8.9), indicating that on low-ambiguity questions raw answers can already produce defensible source grounding. On Legal Precision, Grounded Completeness, and Practical Legal Usefulness the raw polygon equals or marginally exceeds the disambiguated polygon. On low-ambiguity questions, a capable base model is doing the right thing without disambiguation infrastructure to help it.

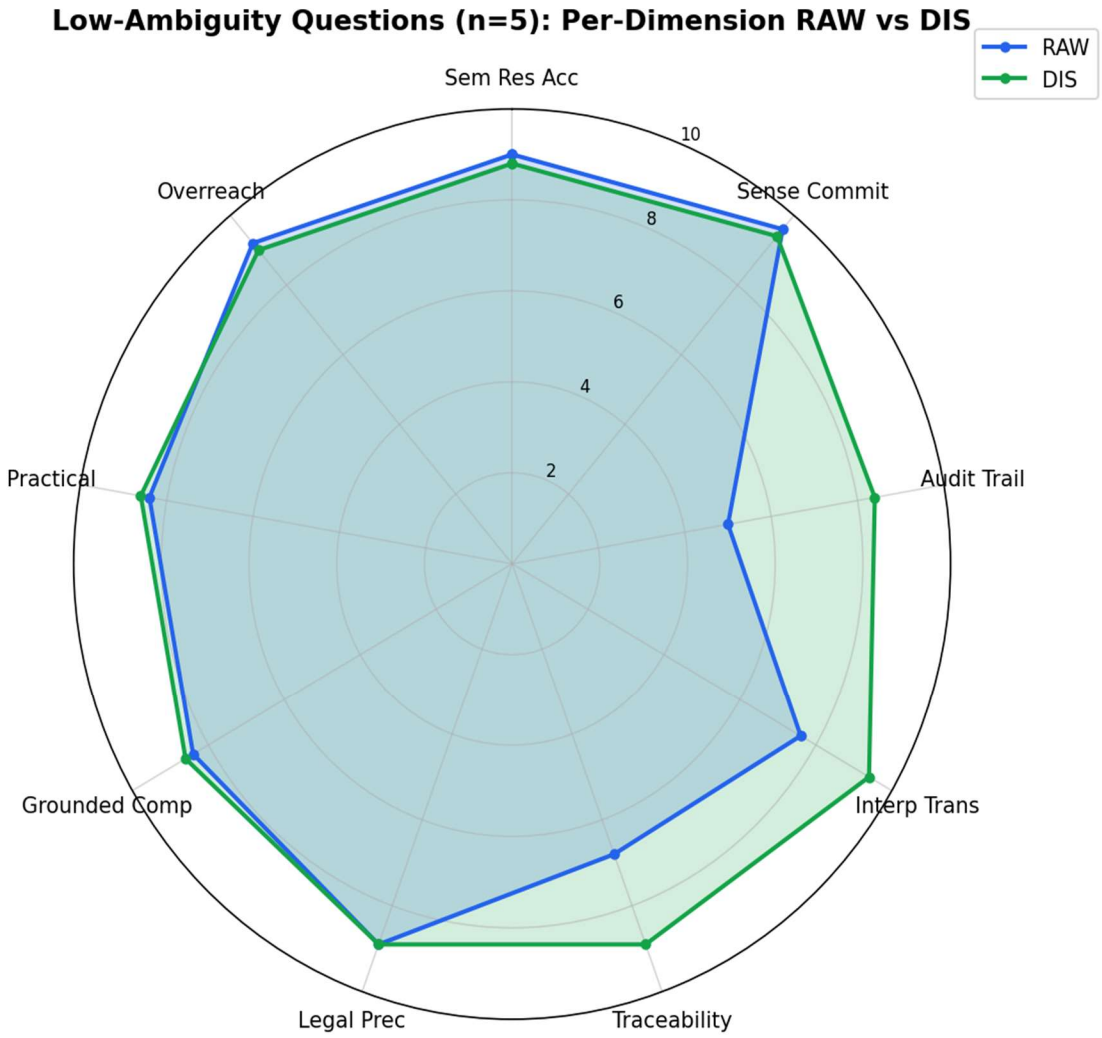

Figure 4. Low-ambiguity questions (n=5): the polygons converge on most dimensions. The two remaining gaps are Traceability and Disambiguation Audit Trail — architectural properties, not writing properties.

The four-chart sequence is the single best illustration of the argument we make in this paper. Disambiguation does not help equally on every question. It helps where help is needed, contributing marginally where it is not — except on Traceability, where the gap is architectural and persists across all tiers. A reader who looks only at the low-ambiguity chart could reasonably conclude that disambiguation is a marginal improvement. A reader who looks only at the high-ambiguity chart could reasonably conclude that it is transformative. Both are correct readings of the same data; the procurement question is which workload the organization actually faces.

11.4 Dimension-by-dimension analysis

Table 3 shows per-dimension mean scores across all 15 questions for each answer stream.

Dimension	RAW mean	DIS mean	Δ

Semantic Resolution Accuracy	7.2	9.3	+2.1
Sense Commitment	9.4	9.7	+0.3
Disambiguation Audit Trail	4.0	8.6	+4.6
Interpretive Transparency	7.1	9.5	+2.3
Traceability & Provenance Integrity	6.1	9.0	+2.8
Legal Precision	8.3	9.0	+0.7
Grounded Completeness	7.1	8.5	+1.4
Practical Legal Usefulness	7.2	8.6	+1.4
Overreach & Hallucination Control	8.8	9.1	+0.4

Three patterns merit discussion.

The largest gap is on Disambiguation Audit Trail (+4.6), with Traceability & Provenance Integrity (+2.8) as the second-largest gap. Raw answers, in the current architecture, do not carry claim-level grounding metadata; they are generated by a single pass that retrieves passages and synthesizes prose without recording per-claim sources, and they do not record the interpretive choice that drove the retrieval. These are not qualities of the model; they are properties of the pipeline. The disambiguated stream, by contrast, records per-claim grounding flags, supporting source IDs, and the full disambiguation audit trail in the recorded artifact. The gap reflects an architectural difference, not a writing difference. Enterprises evaluating legal AI for governance-sensitive workflows should treat these two dimensions as the most important single test: they determine whether claims and interpretive choices can be audited back to source.

Disambiguation-specific dimensions show an uneven pattern across ambiguity severity. Disambiguation Audit Trail (+4.6), Interpretive Transparency (+2.3), and Semantic Resolution Accuracy (+2.1) all show meaningful gaps across the dataset. Sense Commitment shows a much smaller all-question mean gap (+0.3) that masks a clear severity pattern: on high-ambiguity questions Sense Commitment RAW is 9.0 against DIS 10.0 (a 1.0-point gap), but on low- and moderate-ambiguity questions RAW sits at 9.6 against DIS 9.4–9.8 — effectively parity. Raw systems commit to a sense whenever the document context permits it; they commit to a wrong sense when the context is genuinely ambiguous. The all-Q average is a poor summary of this dimension and should be read against the severity-stratified view in Section 11.3.

Legal craft dimensions show modest gaps of 0.7–1.4 points. Legal Precision (+0.7), Grounded Completeness (+1.4), and Practical Legal Usefulness (+1.4) all improve under disambiguation, but modestly. These three dimensions score the answer as a piece of legal writing: terminology, coverage, usability. A capable language model produces reasonable legal-craft output even without disambiguation support. Disambiguation refines rather than transforms the writing-quality dimensions; the architectural value is concentrated in the governance-sensitive dimensions above. Section 11.7 shows how trust-adjusted scoring rescales Practical Legal Usefulness against Traceability — a different lens that magnifies the Practical gap by conditioning it on grounding strength.

11.5 Where raw answers approach disambiguated performance

Not every dimension and not every question shows a substantial gap. Several patterns are worth flagging because they refine the headline finding.

Low-ambiguity questions narrow the gap to under a point on most dimensions. For questions like *"Are there any liens recorded against the collateral?"* — where the term *lien* is unambiguous in the secured-transactions context of the document — the raw answer reaches Grounded Completeness and Legal Precision scores within half a point of the disambiguated answer. The raw model is, in these cases, doing the right thing: it picks the obvious sense without needing the disambiguation infrastructure.

Raw answers occasionally produce more comprehensive coverage. In a small number of cases (notably the material-default and set-off-rights questions), the raw answer enumerates adjacent provisions the disambiguated answer omits as out of scope. Under the rubric's calibration, this can register as a slight Grounded Completeness lift for the raw stream, though usually offset by lower scores on Sense Commitment and Semantic Resolution Accuracy. The pattern reflects a real trade-off: a hedging answer that *covers* multiple interpretations may be more complete in a narrow sense, but is less operationally usable.

Practical Legal Usefulness shows the smallest gap among governance-sensitive dimensions. Even a hedging, multi-sense raw answer can read as practically useful to a reviewer with the time and expertise to extract the relevant sense. The dimension scores the answer from a working reviewer's perspective; it does not penalize the absence of explicit interpretive disclosure as heavily as Audit Trail does. This is an intentional rubric design choice: usability and auditability are related but distinct properties.

11.6 Reading the visual: shape comparison

The four figures above (Figures 1–4) make the per-dimension picture concrete in a way that scoring tables alone do not. The shape difference between the two polygons in Figure 1 is exactly the shape difference between an answer system designed for output quality and one designed for governance. The shape difference *changes* across Figures 2–4 in a way that scoring tables alone cannot show: it is large in the high-ambiguity tier, modest in the moderate tier, and nearly absent (Traceability aside) in the low-ambiguity tier.

This visual reading aligns with the procurement implication: a buyer comparing two legal AI systems on fluency alone will see them as substitutable; a buyer comparing them on the full governance shape sees one as a reliable substrate for regulated workflows and the other as not. The reading also clarifies the trade-off: if the organization's workload is dominated by low-ambiguity questions, the disambiguation infrastructure is overhead more than it is leverage. If the workload contains a meaningful fraction of high-ambiguity questions — and most enterprise legal workflows do — the architecture is operating on the questions that matter most.

11.7 Trust-adjusted scoring: a worked demonstration

The methodology in Section 9 noted that the framework supports an optional trust adjustment — a post-evaluation transformation that down-adjusts dimensions whose

reliability depends on others. The configuration used in the run reported here adjusts Practical Legal Usefulness by Traceability & Provenance Integrity: an answer cannot be more practically useful than its grounding permits. The original scores are preserved; the adjustment is a view-only calculation that lets a reviewer see what the practical score implies once it is conditioned on whether the supporting evidence is actually there. This subsection demonstrates the adjustment on the five questions in the dataset where the rule actually fires — four high-ambiguity questions (Q1, Q8, Q9, Q13) and one low-ambiguity question (Q2), all of which have raw Traceability scores low enough to land in an adjustment band.

Trust adjustment exists to expose a specific failure mode: an answer that reads as fluent and apparently useful but cannot be audited back to source. The dimensions most likely to score well in conventional review — Practical Legal Usefulness, Legal Precision — are also the dimensions most likely to mask grounding failures. The rule used in this run reduces Practical Legal Usefulness when the Traceability & Provenance Integrity score for the same answer falls into a low band, with the magnitude of the deduction depending on both the Traceability band and the question's ambiguity severity. The bands used here are summarised in the table below; UNASSIGNED severity is treated as Low, and adjusted scores are clamped at zero. The rule fires across severity tiers — not only on high-ambiguity questions — whenever the raw Traceability score lands in a band. In this dataset it fires on five of the fifteen questions; their per-question effects are tabulated below, followed by the aggregate visualisation in Figures 5 and 6.

Severity	Traceability score (source)	Deduction to Practical Legal Usefulness
High	0 – 2	−4.0
High	3 – 4	−2.5
High	5 – 6	−1.5
Moderate	0 – 2	−3.0
Moderate	3 – 4	−1.5
Moderate	5 – 6	−0.5
Low	0 – 2	−2.0
Low	3 – 4	−1.0

Per-question effects of the trust-adjustment rule on the five affected rows in this dataset:

Q	Severity	Trace	Practical RAW	Adj	Overall RAW	Adj
Q1	High	3	6.0	3.5	5.4	5.2
Q2	Low	1	8.0	6.0	8.1	7.9
Q8	High	1	2.0	0.0	2.8	2.6
Q9	High	2	8.0	4.0	6.6	6.2
Q13	High	2	2.0	0.0	4.7	4.5
Mean	—	—	5.2	2.7	5.52	5.27

Figure 5 plots the unadjusted view of the five affected questions. The raw polygon (blue) collapses on the architectural dimensions but shows Practical Legal Usefulness at 5.2 — sitting well above the Traceability score (1.8) and only modestly below the Legal Precision score (7.2). A reviewer reading the answers alone might judge them as moderately useful in practice, unaware that the supporting grounding across these five questions is substantially weaker than the practical-usefulness score implies.

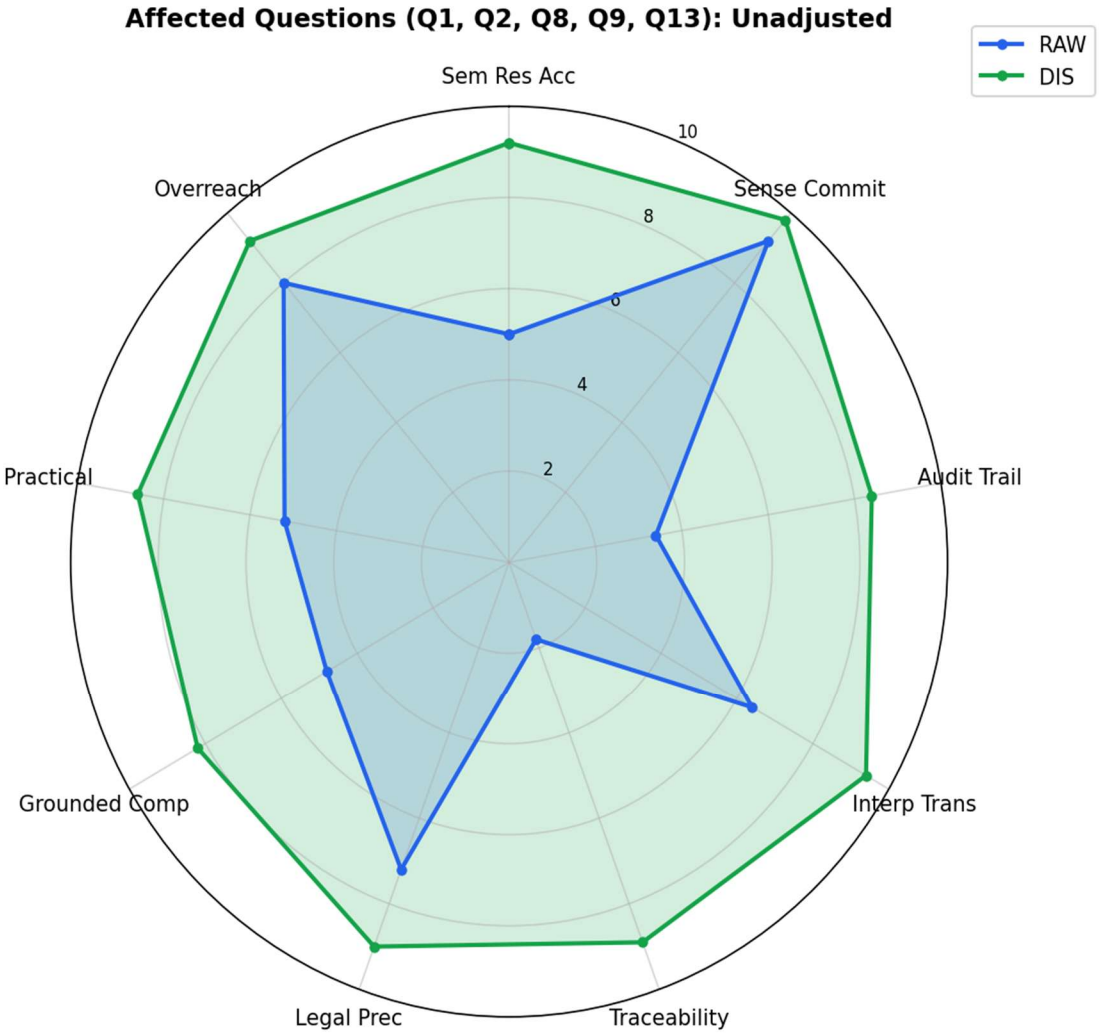

Figure 5. The five affected questions (Q1, Q2, Q8, Q9, Q13), unadjusted view. Raw Practical Legal Usefulness sits at 5.2, masking the underlying Traceability score of 1.8.

Figure 6 plots the same five questions with trust adjustment active. The disambiguated polygon (green) is essentially unchanged — DIS Traceability across these questions is high (mean 8.9), so DIS Practical (8.6) is not pulled down by the rule. The raw polygon (blue) collapses further on the Practical spoke, marked with an orange ring: the adjusted mean is 2.7, down from 5.2 (a 2.5-point drop across the five-question subset). The visual change is dramatic precisely on the spoke a conventional review would have credited most highly. The raw answers' apparent practical usefulness was riding on grounding that was not actually there, and the adjustment exposes it.

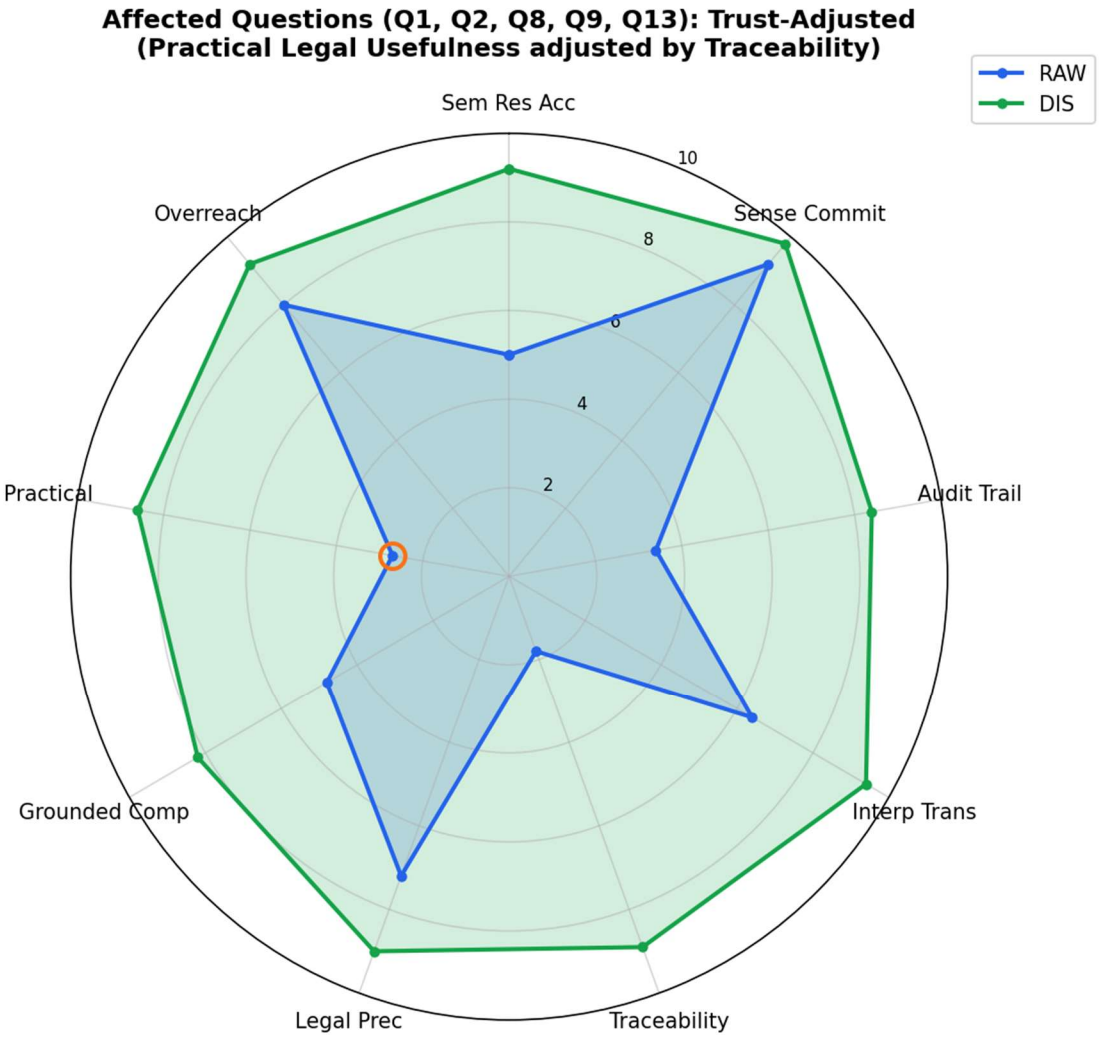

Figure 6. The five affected questions, trust-adjusted view. Raw Practical Legal Usefulness drops from 5.2 to 2.7 once conditioned on Traceability. The DIS polygon is unchanged because DIS Traceability already supports DIS Practical.

The effect on the per-question composite scores is small in absolute terms because Practical Legal Usefulness carries a weight of 1.0 out of a rubric total of 10.0 — about 10% of the overall score. The per-question table above shows the per-Q overall moves: Q1 5.4→5.2, Q2 8.1→7.9, Q8 2.8→2.6, Q9 6.6→6.2, Q13 4.7→4.5 (mean 5.52→5.27). The corresponding DIS overalls are essentially unchanged because DIS Traceability already supports DIS Practical across the subset. The point of trust adjustment is not to dramatically restate the composite. The point is to make visible, on the per-dimension view, the dimension that the system's overall score is implicitly trusting — and to show what happens when that trust is withdrawn.

Two facts about the adjustment are worth noting for procurement. First, the adjustment is configurable: different deploying organizations may declare different "X adjusted by Y" rules to match their own reliability assumptions, and the framework supports declaring them in the rubric. The Practical-by-Traceability rule used here is one defensible choice;

alternatives such as Legal-Precision-by-Audit-Trail or Grounded-Completeness-by-Traceability are equally implementable. Second, the adjustment is most consequential precisely where governance review matters most — on the ambiguous, high-stakes questions where unsubstantiated practical usefulness is operationally dangerous. A system being procured for governance-sensitive workflows should be evaluated both with and without trust adjustment; the gap between the two views is itself a measurement of how much of the system's apparent value depends on its grounding being real.

11.8 LLM-only disambiguation as a control test

A predictable procurement question is whether the outcomes documented in the preceding sections could be achieved more simply — by asking the language model to disambiguate the question before retrieval, without routing through a curated knowledge graph. To address this question directly, the same fifteen questions were run through a control condition we call RAW+: the baseline answering pipeline with a Stage 0 LLM call added to rewrite the question before retrieval.

The RAW+ implementation. Stage 0 is an LLM call that rewrites the user's question to resolve ambiguous terms before retrieval. Four design choices shape how the results read. (1) Stage 0 is document-blind: it uses only the LLM's parametric prior and never sees the source document. (2) Stage 0 has no curated candidate set; it can pick any sense the model's training has encoded. (3) The rewritten query drives BM25 retrieval only. The answer-generation prompt receives the user's original question together with the retrieved passages, not the Stage 0 rewrite. (4) Stage 0 produces no structured trace; its sense commitment is invisible in the answer text and visible only in which passages were retrieved. This is the simplest LLM-only disambiguation baseline — a frontier model asked to pick a sense, without any of the curatorial scaffolding the KG provides.

Overall outcome. On four of the fifteen questions, the Stage 0 rewrite produced retrievals that drove the answer into a frame inapplicable to the document. On the questions where the LLM's parametric prior correctly identified the relevant sense — including most of the low-ambiguity questions and Q1 (where "interest" was correctly disambiguated to Security Interest in a securities-loan context) — RAW+ produced answers substantively equivalent to DIS. A counterintuitive secondary finding emerged: on two questions where the original RAW baseline was already producing acceptable answers, the RAW+ Stage 0 rewrite worsened the answer rather than improving it. The naive LLM-disambiguation pipeline does not just sometimes fail to add value; it sometimes subtracts value. Two of the catastrophic failures are reproduced below in full, with the evaluator's verbatim commentary, because the failure mode is more vivid in the evaluator's words than in aggregate score deltas.

Illustrative case 1: Q12 — "What material representations survive closing?" The question concerns material representations and their survival in the securities loan agreement; the document addresses the survival concept explicitly in Section 21. DIS, with the KG's domain-tagged candidate set, resolved "closing" to a contract-closing sense and constructed an answer engaging with the document's actual survival provisions; DIS scored 8.7 overall. RAW+, with the LLM's document-blind parametric prior, rewrote the query

under a different sense — namely accounting-period close. BM25 retrieved passages weakly matching that frame, and the answer-generation LLM constructed a response within it. The evaluator's verbatim commentary on the resulting RAW+ Semantic Resolution Accuracy score:

> *The answer completely misconstrues the question's context, treating 'closing' as an accounting period close and discussing financial statement line items and disclosures, none of which appear in or relate to the securities loan agreement. An experienced legal reader would interpret 'closing' here, if anything, in transactional/contractual terms (e.g., termination or completion of a loan), and the document explicitly addresses survival of obligations in Section 21, not accounting concepts. The response thus adopts a sense (accounting-period closing) that is inconsistent with the document and sustains that incorrect sense throughout.*

RAW+ scored 1.1 overall on Q12, against DIS's 8.7. Semantic Resolution Accuracy: 0. Grounded Completeness: 0. The answer was internally coherent, fluently written, and about an entirely different document type.

Illustrative case 2: Q14 — "What interest survives default?" The KG considered five candidate senses for "interest" and resolved to Security Interest in the secured-transactions domain — appropriate for the securities loan agreement. DIS produced an answer engaging with the document's actual provisions on post-default remedies (Sections 13, 14, 15, and 21); DIS scored 9.1 overall. RAW+ committed to a sense outside the KG's candidate set entirely: estates-law interest. The evaluator's verbatim commentary:

> *The question clearly refers to contractual or security interests that survive a default in this securities loan agreement, particularly in provisions like Sections 13, 14, 15, and 21 where remedies and rights (including interest on unpaid amounts) survive termination and default. The answer instead shifts entirely to a property/estates-law sense of 'interest' (decedent's estates, fee simple subject to condition subsequent), which is inapplicable to the document's context. This reflects both a wrong sense of the term and a full commitment to that wrong sense throughout the reasoning.*

RAW+ scored 4.0 overall on Q14. The estates-law sense exists in legal English; it was not in the KG's candidate set because the KG's domain tags excluded property-law concepts from a secured-transactions document. The LLM, lacking such domain tags, accessed it from training and committed to it.

What the catastrophes share, and what the KG provides. Both failures share a single mechanism: the document-blind Stage 0 rewrote the query under a sense that exists in general legal English but is inapplicable to the document at hand. BM25 retrieval, keyed to the rewritten query, returned passages that either weakly matched the wrong sense or failed to surface the document's actual operative provisions. The answer-generation LLM, given those passages and the user's original question, constructed an answer in the retrieval's frame. None of these steps individually was unreasonable; the cascade was. The KG-anchored architecture replaces three of these steps with curated alternatives — document-scored candidate selection in place of document-blind parametric-prior

selection; a curated candidate set that filters out senses inapplicable to the document type; and a structured commitment recorded in the trace rather than an invisible commitment encoded only in which passages were retrieved. The gap between RAW+ outcomes and DIS outcomes, including the regression cases where RAW+ scored below the original RAW baseline, is the gap that ontology curation closes. The naive LLM-disambiguation baseline is not a substitute for the architecture; it is the failure mode the architecture is designed to prevent.

12. Key Observations

Several observations from the evaluation generalize beyond the specific dataset.

Interpretation-aware systems reduce semantic drift, not just semantic error. The Sense Commitment and Semantic Resolution Accuracy gaps reflect not only "the right sense was chosen" but "the right sense was *maintained* through the answer." Raw answers frequently pick a sense early and then slip into vocabulary from a rejected alternative later in the same response — a failure mode that is invisible to surface fluency assessment but corrosive to legal review. The disambiguation pipeline, because the chosen sense is baked into the augmented query, produces answers that stay inside one frame.

Traceability is not a feature; it is a precondition for legal review. A claim that cannot be traced to a source passage cannot be reviewed against the source. The 4.6-point Disambiguation Audit Trail gap (paired with the 2.8-point Traceability gap) is the most architecturally consequential finding in the evaluation, because it is not a property the raw system can acquire by writing better prose; it requires structural changes to the answer-generation pipeline. This implies that legal AI systems should be procured on architectural properties (do they record claim-level grounding?), not only on output evaluation.

Auditability improves governance readiness in ways that compound. The disambiguation audit trail — the structured record of which senses were considered, which was selected, and on what evidence — has value beyond the individual answer. It enables consistent review workflows (reviewers know where to look), supports challenge defensibility (the choice is defensible against a counter-reading because the reasoning is recorded), and provides a substrate for regulatory audit (the system's interpretive history is queryable). These benefits accrue over time as the audit corpus grows.

Raw answers tend to overgeneralize under pressure. When a raw model encounters a question with high ambiguity, the typical failure mode is not to refuse or to ask for clarification, but to produce a multi-sense answer that hedges. This is operationally dangerous: an answer that covers multiple legal senses without committing to one is functionally indistinguishable, to a reviewer, from an answer that does not understand the question. A system that says less, more precisely, is more useful than one that says more, less precisely.

The smallest gaps appear where the underlying model is most capable. Base-model legal writing capability has improved substantially across recent model generations, and the modest gaps on Legal Precision and Practical Legal Usefulness reflect this. The implication is not that disambiguation is becoming less relevant; it is that the *kind* of value

disambiguation provides is shifting away from "better prose" and toward "auditable interpretive commitments." Procurement criteria should follow this shift.

13. Operational and Economic Risk

The technical findings reported above translate into measurable operational and economic exposures for organizations deploying legal AI. A system that produces interpretively unstable answers does not merely produce worse answers; it produces answers that consume more downstream resources to review, defend, and correct. This section restates the architectural distinction between traditional retrieval-augmented generation and governance-aware interpretation in the language of operational risk categories that enterprise procurement, legal operations, and risk functions actually track.

The risks fall into seven categories. Each is named in the language of operational risk rather than the language of model evaluation, because that is the language in which deployment decisions are made.

1. Reviewer rework cost. When an AI-assisted legal output is produced under an ambiguous question and the underlying sense choice is not surfaced, the reviewer must reconstruct the interpretation from the output text alone. This is silent work. It does not appear on a timesheet as "disambiguation"; it appears as "review time" — and the time is meaningfully longer than the comparable review of a system that exposes its interpretive commitments. In a high-volume legal operations context (contract review queues, regulatory mapping projects, due diligence batches), reviewer rework compounds across documents. A two-minute interpretive reconstruction per document, across a 5,000-document review, is more than two reviewer-months of unattributed cost.

2. Silent semantic drift. Raw answers tend to commit to a sense early in the output and then drift into vocabulary from a rejected alternative later in the same answer. This drift is invisible to surface fluency assessment — the answer reads well — but it is corrosive to legal review. The reviewer either notices the drift and must reconcile it (rework, see above), or fails to notice and signs off on an answer that internally contradicts itself. Both outcomes are operational liabilities. The first costs time; the second costs defensibility. Governance-aware interpretation does not eliminate drift, but the recorded sense commitment makes drift detectable: a reviewer or downstream check can verify that the answer stays inside the declared sense.

3. Inconsistent legal memos and output variance. Across a portfolio of related questions — for example, twenty contract-interpretation queries against the same family of agreements — a non-disambiguated system can resolve the same polysemous term differently on different occasions, with no audit trail to explain the variance. The downstream artifact is inconsistent legal memos: two outputs that disagree about the meaning of the same operative term, with no recorded reason for the disagreement. Beyond the immediate quality problem, output variance is a procurement liability: it undermines the operational predictability that legal operations functions rely on to plan capacity, scope projects, and price work to clients.

4. Audit failure risk. Audits — internal compliance audits, external regulatory audits, professional-conduct reviews — require the audited party to produce a record of the reasoning behind a decision, not only the decision itself. A legal AI system that produces only answers, without a recorded interpretive history, places the burden of audit-trail reconstruction on the human user. In practice this means either retrofit documentation (slow and partial) or audit findings indicating that the AI-assisted workflow cannot be reconstructed (the more serious outcome). Governance-aware interpretation produces the audit artifact as a normal byproduct of operation, which converts an audit-failure exposure into a documented compliance posture.

5. Regulatory explainability obligations. The regulatory trajectory across major jurisdictions — the EU AI Act for high-risk AI systems, sectoral guidance from financial and prudential regulators, attorney professional-conduct revisions on generative AI use — points consistently toward documented explainability. Systems deployed today will be measured tomorrow against explainability standards that were not in force at procurement. A traceable interpretive architecture is a hedge against this regulatory drift: the cost of producing the trace artifact is fixed and amortized across all queries; the cost of retrofitting explainability onto an opaque system is variable, episodic, and potentially compounding across years of accumulated outputs.

6. Inability to defend AI-assisted workflows. In matters where AI-assisted analysis is challenged — adverse counsel in litigation, regulatory examination, professional-conduct complaint, internal escalation — the defending party must be able to reconstruct what the system did, what it considered, and what evidence supported each material claim. Without that record, the defense reduces to an attestation that the output was reviewed and looked reasonable, which is an exposed position. The governance-aware trace converts the defensive posture from attestation to documentation. The economic value of that conversion is impossible to quantify in advance of an actual challenge, but it is large in the cases where it matters.

7. Escalation burden. A non-disambiguated system that produces an interpretively wrong answer typically does not flag the error; the error must be detected by a human reviewer. Detection rates depend on reviewer expertise, fatigue, and time pressure. Each detected error triggers an escalation back to the original requester, into a clarification loop, sometimes into a rework cycle that consumes more time than the original query. Each undetected error becomes a downstream correction cost: memos rescinded, advice withdrawn, decisions revisited. Governance-aware interpretation reduces escalation burden in two ways — by surfacing low-confidence interpretive selections for review at the moment of generation (cheaper to escalate early), and by recording the audit trail in a form that supports rapid post-hoc verification (cheaper to escalate late).

These categories overlap; few real-world incidents fall cleanly into only one. An audit finding may also be a litigation exposure; a reviewer-rework problem may also be an output-variance problem. The point is not to provide a precise risk-by-risk accounting but to make explicit what is at stake operationally when a legal AI architecture is chosen. The technical evaluation in earlier sections measures the system's properties; this section names the consequences of those properties for the organizations that deploy it.

A note on quantification. None of the figures suggested above (reviewer-months, escalation rates, audit findings) are derived from a controlled study. They are illustrative of the risk categories rather than measurements of them. A defensible quantification of operational risk reduction from governance-aware interpretation would require longitudinal deployment data — review-time deltas, escalation-rate comparisons, audit-outcome comparisons across matched workloads — which is exactly the kind of evidence that procurement decisions today are typically made without. The architectural argument does not require precise quantification to be material: a fixed-cost trace architecture against a set of variable-cost downstream exposures is a defensible procurement posture even before the deltas are measured.

14. Enterprise Implications

For organizations evaluating legal AI deployment, the findings translate into specific procurement and operational implications.

Law firms. For matter-relevant questions where ambiguity is structurally high — contract interpretation, regulatory mapping, multi-jurisdictional comparisons — the disambiguation lift is large enough to be operationally material. Firms should evaluate legal AI not only on representative output samples but on the system's ability to surface its interpretive commitments and trace its claims. A system that produces high-quality answers without an audit trail will create downstream defensibility problems even when individual outputs are correct.

Enterprise legal operations. Repeated workflows (contract review queues, due diligence batches, regulatory mapping) compound the benefit of auditability. A traceable system supports consistent review patterns and creates a corpus of interpretive choices that the operations team can examine for drift, bias, or systematic error. A non-traceable system requires every workflow batch to be reviewed afresh, with no compounding learning over time.

Compliance and AI governance. The traceability requirements emerging in regulatory frameworks (the EU AI Act, sectoral AI guidance from bank regulators, attorney professional-conduct rules in jurisdictions revising guidance on generative AI use) generally require some form of documented interpretive process and source attribution. A governance-aware legal AI architecture is closer to satisfying these requirements out of the box than a non-disambiguated system that must retrofit the documentation. Procurement decisions made today should factor in the regulatory trajectory, not just the current letter of the law.

Litigation support. In matters where AI-assisted research or analysis may be challenged on cross-examination or in discovery, the audit trail is a defensive asset. A reviewer who can show the interpretive choices the system made, the alternatives it considered, and the sources supporting each claim has a substantially stronger position than one who can only attest to having reviewed the system's output.

Regulated industries. Banking, insurance, healthcare, and similar sectors where legal AI is being deployed under heightened scrutiny benefit disproportionately from interpretive

transparency. The cost of a defensible audit trail is fixed; the cost of being unable to produce one is variable and potentially substantial.

A note on procurement framing. The conventional procurement question for legal AI — "does this system produce good answers?" — is necessary but not sufficient. A more complete framing is: *does this system produce answers I can defend?* The two questions are related but distinct. The disambiguation lift documented here addresses the second question more directly than the first.

15. Limitations

The evaluation reported here has several constraints that should temper how strongly its conclusions are read.

Sample size. Fifteen questions against a single source document is a small sample by any benchmark standard. The findings should be read as patterns suggestive of a more general phenomenon, not as statistically conclusive measurement. We make claims about the *direction* and *severity-correlation* of the effects with high confidence; claims about the precise magnitude of per-dimension lifts should be treated as approximate.

Single document. All questions were evaluated against one securities loan agreement. The document is representative of commercial finance contracts but does not span the full range of legal domains (litigation, regulatory, criminal, IP, employment, M&A, real estate). The disambiguation lift on a complex M&A agreement might differ in magnitude from the lift on a loan agreement; we have no data to make that claim either way.

Evaluator subjectivity. Scores are produced by an LLM evaluator applying rubric scoring prompts to (document, question, answer) triples. The evaluator is itself a probabilistic system; identical inputs can produce slightly different scores across calls. We have not run a formal inter-rater reliability study, and the reported scores should be understood to include some evaluator noise. The patterns we report are large enough that we believe they survive that noise, but we cannot quantify the residual uncertainty.

Methodology. The framework is deliberately not an academic benchmark. Calibration anchors in the rubric reflect the framework author's judgment about what each score band should represent. A different rubric author would calibrate differently. The framework's value is in *consistency across compared systems* rather than in absolute score values.

No human reviewer baseline. Scores are produced by the evaluator LLM, not by panels of legal experts. A human-reviewed comparison would be more authoritative on dimensions like Legal Precision and Practical Legal Usefulness, where domain judgment is most consequential. We treat the LLM evaluator as a reasonable proxy for an experienced legal reviewer applying the calibration anchors, but this assumption deserves empirical validation.

16. Future Work

Several directions would strengthen the evidentiary base.

Larger datasets. Extending the evaluation to several hundred questions across diverse document types — credit agreements, M&A contracts, regulatory filings, court opinions — would substantially increase confidence in the generalizability of the patterns reported here.

Multiple legal domains. Repeating the framework against questions drawn from litigation, regulatory compliance, IP, employment, and corporate transactions would clarify whether the severity-correlated lift pattern generalizes across legal practice areas.

Human evaluator panels. A study comparing LLM-evaluator scores to scores produced by experienced legal reviewers would validate (or correct) the calibration of the rubric anchors and provide a more authoritative read on dimensions where domain judgment matters most.

Benchmark automation. A reproducible benchmark suite — fixed document corpus, fixed question pool, fixed rubric — would allow comparative measurement across legal AI vendors. The framework's design (snapshot-based rubric storage, deterministic question imports, per-cell rerun support) is consistent with this goal, but the benchmark suite itself would require a coordinated multi-stakeholder effort.

Probabilistic interpretation scoring. The current rubric treats sense selection as a discrete choice. A more sophisticated framework would score the probability distribution the system assigned to candidate senses, rewarding systems that surface calibrated uncertainty rather than (or in addition to) systems that commit confidently to a single choice. This direction is consistent with regulatory trajectories that emphasize uncertainty disclosure.

Ontology-driven governance. The disambiguation pipeline depends on a legal ontology that enumerates candidate senses per term. The quality of the ontology bounds the quality of the disambiguation. Future work should investigate how to evaluate the ontology itself: coverage, granularity, domain partition. An ontology-aware governance framework would be able to attest not only that the system disambiguates, but that the disambiguation space against which it does so is appropriate to the matter being analyzed.

Trust-rule calibration studies. The framework supports trust-adjustment rules that down-adjust dependent dimensions based on upstream reliability scores. The calibration of these rules — when to adjust, by how much, against which dependencies — is currently set by the rubric author. Studying how trust adjustments correlate with reviewer disposition (do legal reviewers actually de-rate answers whose traceability is weak?) would put the trust-adjustment mechanism on firmer empirical ground.

17. Conclusion

Legal AI systems have improved substantially as language models. They have improved less as governance objects. A typical contemporary system produces fluent, often correct legal prose; the same system frequently cannot explain which interpretation it committed to, why it committed to it, or what evidence supports the resulting claims. For consumer applications this is a quality concern. For enterprise legal deployment — where outputs

feed into defensible memos, regulated workflows, and decisions whose audit trail may be examined years later — this is a governance concern.

The evaluation reported here documents the size of the gap between two architectural patterns: a raw answer pipeline and a disambiguation-driven, provenance-aware pipeline. The lift averages roughly two points on a ten-point composite, but the more informative statistic is the severity correlation: high-ambiguity questions show a 4.5-point gap, low-ambiguity questions show under one. The pattern is consistent with the architecture doing what it claims to do — helping where help is most needed.

We do not claim that disambiguation solves legal AI, that raw systems are unusable, or that interpretation-aware systems are universally superior. The evidence is more specific. Governance-aware semantic interpretation materially improves reliability and operational trustworthiness in ambiguity-sensitive legal workflows. The improvement is largest on the dimensions that matter most for legal review and regulatory defensibility — interpretive transparency, audit trail, and claim-level provenance. It is more modest on dimensions where capable base models already perform well — legal precision, practical usefulness, completeness of coverage.

For organizations procuring or deploying legal AI, the operational implication is straightforward. Output quality is necessary but no longer sufficient. The relevant procurement question is whether the system's outputs are defensible — whether the interpretive choices it makes are visible, the evidence it relies on is traceable, and the reasoning it produced can be reproduced under review. Systems that meet this bar are operationally trustworthy in ways that systems judged on fluency alone are not. The architectural pattern that enables it — explicit ambiguity detection, recorded sense selection, interpretation-aware retrieval, claim-level provenance — is the substrate on which the next generation of regulated legal AI will need to be built.

Document version 1.0. Prepared from real evaluation data (15 questions against a single securities loan agreement). See Section 15 for limitations.