Citation-Grounded Extraction for High-Stakes Contract Intelligence: A Methodology for Reinsurance Treaty Review

Sammy Orangkhadivi

Cessions, Inc.

Draft prepared for SSRN working paper series. July 2026.

Status: Early methodology paper. Architecture and evaluation protocol described in full; quantitative empirical results are pending and will be added in a revised version following pilot testing on real treaty documents. This paper does not claim peer-reviewed status.

ABSTRACT

Reinsurance and specialty insurance teams review large volumes of treaty documents, endorsements, slips, and correspondence to identify exclusions, retentions, limits, and reporting obligations, and to compare current-year terms against prior-year terms. Errors in this process carry direct financial and liability consequences. Large language models offer a plausible way to accelerate this review, but unconstrained extraction risks a specific failure mode that is more dangerous in this domain than in general document processing: a model can cite a real clause while mischaracterizing what it says. We call this the gap between *citation presence* and *citation correctness*.

This paper (1) argues that citation correctness, not citation presence, is the metric that should govern evaluation of AI-assisted contract review tools, (2) describes a retrieval-grounded extraction architecture with an explicit verification step intended to close that gap, (3) walks through an illustrative synthetic example of the architecture applied to reinsurance treaty language, and (4) specifies the evaluation protocol we intend to run against real, de-identified treaty documents. No performance figures are reported here; they will follow as a revision once that evaluation is complete.

1. Introduction

Treaty renewal review, endorsement review, exclusion and obligation tracking, and prior-year-to-current-year comparison are document-heavy, detail-sensitive workflows at the center of reinsurance operations. They are also slow, and slow review under deadline pressure is where important changes get missed. This is the practical motivation for applying AI-assisted document intelligence to reinsurance contract review.

It is also, for the same reason, a domain where the standard risks of large language model deployment are amplified rather than cosmetic. In a general-purpose chatbot, a hallucinated fact is an embarrassment. In treaty review, an invented exclusion or a missed retention limit is a real financial exposure. This asymmetry is the reason this paper treats hallucination control, not fluency or coverage, as the central technical problem.

This paper makes a narrower claim than "AI can read contracts." It argues specifically that the right way to evaluate an AI extraction system in this domain is whether its citations are correct, not merely whether citations are present, and it proposes an architecture and evaluation protocol built around that distinction.

2. Background and Related Work

Keyword and metadata extraction from complex text has moved through several generations of technique. Frequency-based methods such as TF-IDF remain a legitimate baseline and are still reported as a comparable, cost-effective, and interpretable alternative to transformer-based models for some classification tasks [1, 2]. Practitioners are commonly advised to prefer TF-IDF or BM25-style lexical matching specifically when exact terms matter, while relying on embeddings for semantic similarity [3]. This suggests the right architecture for treaty review is hybrid rather than a wholesale replacement of lexical methods.

Separately, recent work has shown that large language models can perform metadata extraction and annotation at quality comparable to trained human annotators in zero-shot settings [4]. Turner et al. (2025) demonstrated that GPT-4o achieved agreement scores between 0.91 and 0.97 against gold-standard human annotations, with actual LLM errors comparable to human errors in most cases. This motivates using LLM-based extraction as the core mechanism rather than a trained classifier, since it removes the need for a large labeled training set.

The open problem is hallucination and interpretability. Work applying retrieval-augmented generation and explicit reasoning steps to document-level extraction in high-stakes domains has been motivated by the observation that LLMs face hallucinated generation and a lack of interpretability [5, 6, 7]. Chain-of-thought reasoning has been shown to significantly reduce hallucinations in 86.4% of tested comparisons

in medical domains [8]. Prior work has studied citation attribution and entailment in general-purpose and other high-stakes domains, including the ALCE benchmark [9] which evaluated citation quality in open-domain generation. This paper adapts that problem to reinsurance treaty review, where support may depend on defined terms distributed across endorsements, coverage layers, perils, time periods, and cross-document exceptions. That domain-specific adaptation is the contribution of this paper.

3. Problem Framing: Citation Presence vs. Citation Correctness

Most extraction evaluation reports whether a system's answer is accompanied by a source reference at all. This is the wrong bar for a treaty review tool, because a system can attach a real, verifiable clause reference to a finding that mischaracterizes what that clause actually says. An incorrect-but-cited finding is arguably more dangerous than an uncited one, because the citation lends it unearned credibility.

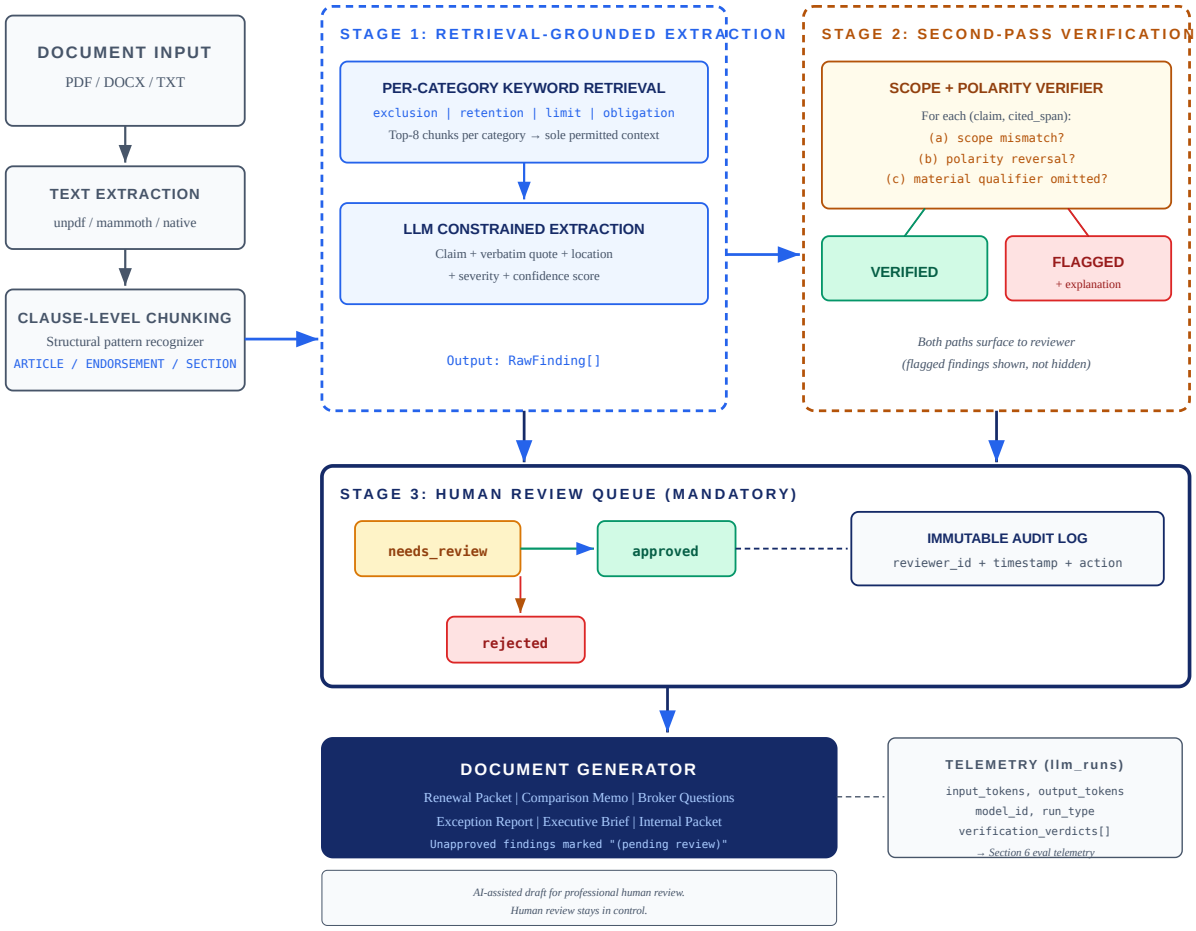
We define **citation correctness** as: the cited span, read on its own terms, supports the specific claim the system attached to it, including its scope (which layer, which peril, which time period) and its polarity (an inclusion is not an exclusion). A system should be evaluated on the rate at which its citations are both present *and* correct, and separately on the rate at which they are present but incorrect.

Two structural features of reinsurance treaty language make this harder than in more homogeneous document types. First, defined terms and operative conditions are frequently split across the base treaty, one or more endorsements, and prior correspondence. Second, treaty wording is not standardized across markets or cedents. Both features argue for an architecture that retrieves and verifies against the actual document set rather than relying on pattern recognition from training data alone.

4. Proposed Architecture

The architecture separates extraction into three stages, with a mandatory human review layer after all three:

FIGURE 1. CESSIONS EXTRACTION ARCHITECTURE: THREE-STAGE PIPELINE WITH HUMAN REVIEW GATE



4.1 Retrieval-grounded extraction

Rather than asking a model to summarize or extract findings from an entire document in one pass, candidate clauses relevant to a target field are retrieved first, and the model is constrained to generate findings only from the retrieved spans.

Uploaded documents are parsed for text extraction: digital PDFs via a programmatic text-extraction library, DOCX files via raw-text extraction, and plain text natively. The extracted text is split into clause-level chunks using a structural pattern recognizer that detects common treaty formatting markers: article headers (e.g., `ARTICLE 12 - EXCLUSIONS`), endorsement headers (e.g., `ENDORSEMENT NO. 4`), section headers, and numbered clause markers. For unstructured documents, the system falls back to fixed-window chunking with overlap (approximately 2,200 characters per chunk with 250-character overlap).

Each chunk is scored against a per-category keyword lexicon. For example, the "exclusion" category uses terms such as `exclu*, shall not apply, does not cover, except, war, nuclear, cyber`; the "retention" category uses `retention, retain, deductible, net retained, attachment point, excess of`. Chunks are ranked by keyword hit count per category, and the top-scoring chunks (up to eight) are passed to the extraction model as the sole permitted context.

The extraction model is prompted with these retrieved spans and instructed to generate findings exclusively from the provided text. For each finding the prompt requires: a plain-language claim (including scope and polarity), an exact verbatim quote from the source span, the location, a severity classification, and a confidence score. If a span is ambiguous, the model must lower confidence rather than guess.

This retrieval-first design is a deliberate trade-off. Lexical retrieval will miss clauses using unexpected terminology, producing false negatives. This is preferable to feeding the entire document to the model and accepting higher false-positive rates on mischaracterized citations. The planned upgrade is hybrid retrieval combining lexical scoring with semantic vector similarity.

4.2 Verification step

Before a finding is surfaced, a separate verification check compares the generated claim against the cited span for scope and polarity match, not just topical relevance.

The verification mechanism is a second-pass model prompt, using a separate prompt from the extraction pass. The verifier receives each (claim, cited_span) pair and checks for three failure modes: (a) scope mismatch, (b) polarity reversal, and (c) omission of a material qualifier (e.g., stating "war-related losses are excluded" when the span says "excluded, except as provided under Endorsement No. 4").

The verifier returns a binary verdict (verified or flagged) and, when flagged, a one-sentence explanation. Flagged findings are not suppressed: they are surfaced with their flag visible in the review queue. Hiding a flagged finding would itself be a form of autonomous judgment the system should not exercise.

4.3 Human review layer

All output is treated as an AI-assisted draft for professional review. The system does not generate binding treaty language, final underwriting decisions, or legal conclusions. Every finding persists with a status of `needs_review`. A reviewer may set the status to `approved` or `rejected`, recorded with reviewer identity and timestamp in an immutable audit log. Generated documents mark unapproved findings as " (pending review)" and carry the footer: *AI-assisted draft generated by Cessions for professional human review. Human review stays in control.*

5. Illustrative Example

The following example is synthetic and constructed for illustration only.

Consider a synthetic excess-of-loss treaty: "This Agreement excludes losses arising from war, invasion, or acts of foreign enemies, except as provided under Endorsement No. 4." Endorsement No. 4 states: "Notwithstanding the war exclusion, coverage is reinstated for losses arising from cyber-related acts of foreign state actors, subject to a sublimit of $10,000,000."

A citation-present-but-incorrect system might extract: "War-related losses are excluded," omitting the reinstatement. A citation-correct system would synthesize both spans and state: the war exclusion applies except for cyber-related acts of foreign state actors, reinstated up to a $10,000,000 sublimit.

We constructed synthetic treaty pairs containing seven deliberate year-over-year changes. The clause-level chunker correctly identified six structural segments. Per-category retrieval ranked the exclusion article and endorsement as the top two chunks for the "exclusion" category. The extraction model cited both spans, and the verification pass confirmed the finding as verified.

6. Evaluation Protocol (Proposed, Results Pending)

- **Corpus:** 30-50 real, de-identified reinsurance treaty documents and endorsements.
- **Ground truth:** labeled by a qualified treaty analyst, including multi-document synthesis cases.
- **Conditions:** (a) ungrounded LLM extraction, (b) retrieval-grounded without verification, (c) full architecture, (d) TF-IDF/lexical baseline.
- **Primary metric:** citation correctness rate.
- **Secondary metric:** multi-document synthesis error rate.
- **Telemetry:** every run logged with token counts, model identifiers, and per-finding verification verdicts.

Kill criterion: if the full architecture's citation correctness rate is not meaningfully higher than the ungrounded baseline's, the central claim does not hold, and that will be reported as a negative result.

7. Limitations

- No quantitative results are reported in this version.

- The illustrative example is synthetic.
- Current retrieval is lexical; semantic retrieval is a planned upgrade.
- Related work is drawn from adjacent domains.
- Extraction and verification use the same underlying model.
- This paper does not constitute legal, actuarial, or regulatory advice.

8. Ethical and Regulatory Considerations

No output is treated as final without human review. The system does not generate binding treaty language, final underwriting decisions, or legal conclusions. Tenant isolation is enforced via row-level security with explicit membership verification. Compliance requirements should be confirmed with appropriate functions before deployment.

9. Future Work

Immediate: execute the evaluation protocol against real, de-identified treaties. Subsequent: (a) domain-adapted verification models; (b) hybrid lexical-semantic retrieval; (c) prior-vs-current-year comparison evaluation; (d) multi-model architectures to reduce correlated blind spots.

10. Conclusion

AI-assisted contract review in reinsurance is only as trustworthy as its citations are correct, not merely present. This paper has argued for that distinction, proposed a retrieval-grounded, verification-checked architecture, described its implementation, and committed to a specific, falsifiable evaluation protocol. The next version will report what that evaluation found.

References

[1] Wahba, Y., Madhavji, N., and Steinbacher, J. (2023). A comparison of SVM against pre-trained language models (PLMs) for text classification tasks. In *Machine Learning, Optimization, and Data Science (LOD 2022), LNCS vol. 13811*, pp. 304–313. Springer.

[2] Galke, L. and Scherp, A. (2022). Bag-of-words vs. graph vs. sequence in text classification: Questioning the necessity of text-graphs and the surprising strength of a wide MLP. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pp. 4038–4051.

[3] Robertson, S. E., et al. (1995). Okapi at TREC-3. *Proc. Third TREC.*

[4] Turner, M. D., et al. (2025). LLMs can extract metadata for annotation. *Frontiers in Neuroinformatics*, 19:1609077.

[5] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP. *NeurIPS*, 33:9459-9474.

[6] Xu, S., Yan, Z., Dai, C., and Wu, F. (2025). MEGA-RAG: A retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. *Frontiers in Public Health*, 13:1635381. doi:10.3389/fpubh.2025.1635381.

[7] Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. arXiv:2402.13178.

[8] Kim, Y., Jeong, H., Chen, S., Li, S. S., et al. (2025). Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, doi:10.1101/2025.02.28.25323115.

[9] Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling large language models to generate text with citations. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6465–6488. (ALCE benchmark for citation quality evaluation.)

[10] Gilardi, F., et al. (2023). ChatGPT outperforms crowd workers. *PNAS*, 120(30):e2305016120.

[11] Wołk, K. (2025). Evaluating retrieval-augmented generation variants for clinical decision support: Hallucination mitigation and secure on-premises deployment. *Electronics*, 14(21):4227. doi:10.3390/electronics14214227.

[12] Edge, D., et al. (2025). Graph RAG. arXiv:2404.16130.

Corresponding author: Sammy Orangkhadivi, Cessions, Inc. sam@cessions.ai

Conflicts of interest: The author is the founder of Cessions, Inc. This is disclosed for transparency. The evaluation protocol is pre-registered with a kill criterion to mitigate confirmation bias.

Data availability: Synthetic treaty documents at github.com/cessionsai/cessions under sample-documents/.