

Authority-Aware GraphRAG for Explainable Tax-Risk Evidence Retrieval: A Controlled Evaluation of Typed Legal Graph Expansion

Thomas Gupta^{1,*}, Arjun Henderson²

¹School of Information Studies, Syracuse University, Syracuse, USA

²Oracle Corporation, Austin, USA

*Corresponding email: t.gupta2823@outlook.com

Abstract: Tax-risk analysis needs evidence sets that include a governing provision, companion rules, and a visible relationship path, rather than a single high-scoring passage. We study the retrieval stage of graph-grounded tax analysis through an executable, controlled benchmark. Building on citation-graph expansion used in tax-oriented retrieval systems, we propose Typed Evidence-Budgeted GraphRAG (TEB-GraphRAG), which jointly scores lexical seeds, relation types, and a deterministic sparse two-topic gate while preserving graph-path provenance. A transparent synthetic collection contains 59 real tax-section identifiers with researcher-authored synopses, 238 researcher-authored directed relations, and 700 templated issue descriptions; 500 held-out queries are scored without learned retriever training. On this benchmark, our method reaches 85.40% Recall@5, versus 69.27% for unrestricted one-hop graph expansion and 59.93% for BM25. The 16.13-point advantage over graph expansion has a topic-cluster bootstrap 95% interval of [11.37, 21.60] points. Under type-label permutation on a balanced 200-case stress subset, mean Recall@5 drops to 0.597; an unrelated-topic distractor reduces full-test recall to 0.636. Typed retrieval therefore depends strongly on metadata fidelity and query focus. These controlled findings do not establish legal correctness, tax-risk classification, or LLM answer quality.

Keywords: GraphRAG, tax-law retrieval, knowledge graphs, explainable AI, provenance, evidence selection, synthetic benchmark.

1. Introduction

A plausible tax position cannot be justified by the surface similarity between a case description and one tax provision. A reviewer may need to inspect a primary provision, a rule governing its application, and adjacent limitations before determining which issues require further inquiry. Retrieval-augmented generation (RAG) makes external evidence accessible to language models [1], but its evidentiary value depends on what the retriever actually supplies. Standard lexical ranking such as BM25 [2] can find a topical provision without collecting the connected material needed to make an explanation auditable.

Graph indexing provides a route beyond isolated text matching. GraphRAG approaches have shown the value of organizing retrieved evidence through structured relationships [3]. Yet a tax graph may contain both legally motivated cross-references and loose topical connections; indiscriminate expansion can use a small context budget on the latter. Legal explanation scholarship further emphasizes the need to distinguish a traceable justification from a plausible narrative [4]. We therefore focus on a narrowly measurable interface: selecting and recording authority candidates before an analyst or a generative model makes any substantive tax judgment.

This paper asks whether typed graph expansion can improve the completeness of a compact tax-evidence packet while retaining mechanically auditable paths. Its testable contributions are: (i) a budgeted relation-aware reranker that retains a strong lexical seed and promotes relevant connected documents; (ii) an optional two-topic routing term that models computational specialization without relying on a

large neural model; and (iii) a transparent controlled evaluation including relation corruption, a source-inspired graph-expansion comparator, ablations, structural negative controls, lexical distractors, and a reproducible audit trail. We delimit conclusions to retrieval; high-stakes legal validity remains a separate research question.

2. Related Research

A. Legal retrieval and domain adaptation

LEGAL-BERT demonstrates how pretraining strategy matters for legal language understanding [5], while LexGLUE [6] and LegalBench [7] establish broad evaluation settings for legal comprehension and reasoning. Neither benchmark should be treated as a tax-expert adjudication set for uncertain tax positions. For case retrieval, BERT-PLI [8] models paragraph-level interactions, illustrating how legal relevance can depend on relations spanning parts of a document. These methods motivate domain-sensitive evidence selection, but they are not trained or compared as neural baselines here.

B. Evidence retrieval and graph structure

Dense passage retrieval [9] and Sentence-BERT [10] offer alternatives to lexical search when query and passage vocabularies differ. On the graph side, KG²RAG uses graph-guided chunk expansion and organization [11], HippoRAG couples a knowledge graph with Personalized PageRank for long-term memory [12], and RAPTOR organizes passages hierarchically [13]. The present study instead operates on explicit tax-section identifiers and typed one-hop edges; it is designed to test whether a five-item evidence budget benefits from structural filtering rather than broad knowledge integration. Execution-guided retrieval and reuse in machine-

learning pipeline generation provides an instructive example of using external evidence alongside direct validation of proposed outputs [26].

C. Attribution, adaptive reasoning, and efficient modules

Self-RAG [14] and corrective RAG [15] address the quality of model-grounded responses and the retrieval conditions under which they are produced. RARR [16] studies post hoc evidence attribution, and FActScore [17] evaluates supported atomic claims in generated text. Here no fluent answer is generated; our path-validity checks evaluate graph structure, not correctness of generated assertions. ReAct [18] frames reasoning as interaction with tools and external information; we regard the evidence packet as a potential input to such a downstream workflow.

Sparse neural experts were introduced by Shazeer et al. [19] and scaled through Switch Transformers [20]. Parameter-efficient adaptation such as LoRA [21] can make legal-domain neural models more economical. Our two-topic gate is a deterministic lexical scorer, not a trained neural mixture-of-experts or a parameter-efficiently tuned LLM. Similarly, GPU vector indexing with FAISS [22] and rank fusion [23] are relevant engineering alternatives but were not used in these runs. A broader RAG taxonomy [24] provides context for separating retrieval-stage evaluation from generation-stage assessment.

3. Method

A. Executable foundation and research question

LLaMA-UTP by Tan et al. incorporates tax-authority nodes and relationships into retrieval expansion and explanation paths [25, Sec. III-E]. We implement the concrete retrieval operation—select initial legal-document seeds, traverse a depth-limited tax graph, and add connected authorities—as the unrestricted graph-expansion comparator. We then modify the expansion scoring rule to account for relation type and a

strict evidence budget. This isolates a transferable technical component from the larger model architecture and allows the graph mechanism to be tested independently of the domain-adapted LLaMA model.

Let q be a synthetic issue description, V an authority registry, and E a set of directed typed edges. A retrieval method returns an ordered packet $P(q)$ of at most $B=5$ document identifiers. Each expanded identifier may carry a provenance record (seed, edge label, destination). The main evaluation target is the fraction of query-specific gold identifiers $G(q)$ recovered inside $P(q)$, rather than prediction of High/Medium/Low risk.

B. Corpus registry, edge types, and provenance

Twenty tax-issue groups were authored around U.S. section designators, including research credits, transfer pricing, interest limitations, foreign tax credits, income timing, and loss carryovers. The 59 distinct identifiers combine Internal Revenue Code and Treasury Regulation section names; some sections occur in more than one issue group. Short explanatory synopses were written by the researchers and must not be confused with authoritative statutory or regulatory text. A directed graph of 238 edges includes the designed labels interprets, companion, and mentions. These are scenario-design associations, rather than extracted or independently verified legal citation relationships. The distinction is consequential: a verified structural path through this graph is not proof that a legal interpretation is valid.

A registry key preserves canonical identifiers (for example, IRC-901, IRC-904, and CFR-1.901-2). In an operational deployment, these would be resolved to versioned primary authority text with a document date, jurisdiction, and citation offsets. This controlled implementation instead returns registry-bound identifiers and paths, which are readily checked against the released JSON graph. Source passage quoting and time-dependent legal authority hierarchy are outside the experimental scope.

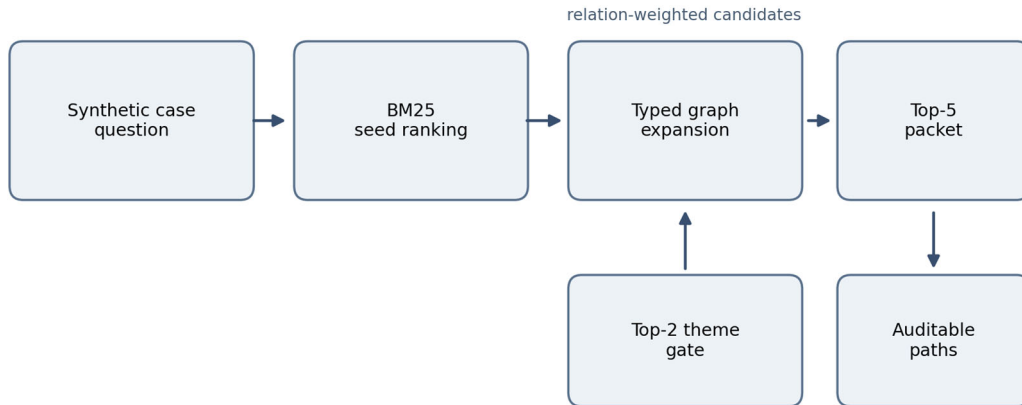


Figure 1. Implemented retrieval pipeline. The candidate packet is compatible with downstream language-model prompting, but this experiment executes retrieval and path assembly only.

C. Baseline retrieval and expansion

We tokenize text into lowercased alphabetic and numeric terms and compute BM25 with $k_1=1.2$ and $b=0.75$ over all 59 synopses. TF-IDF with unigram-bigram features provides a second purely lexical comparator. For the source-inspired graph method, the top three BM25 seeds are used for one-hop outgoing traversal without distinguishing relation categories. Expanded candidates receive a combination of normalized

BM25 relevance and seed strength; the five highest-scoring candidates are emitted. The implementation follows the graph-expansion operation rather than the original paper’s entire neural stack, specialized dense embeddings, or language-model training.

D. Typed evidence-budgeted reranking

Our reranker starts from the same top-three BM25 seeds. For each reachable candidate d from seed s with edge type t ,

the score is $S(d|s,q)=0.25 b(d,q)+0.83 b(s,q) w(t)+0.12 a(d,q)$, where b is BM25 normalized by the largest query score. Typed weights are $w(\text{interprets})=1.00$, $w(\text{companion})=0.90$, and $w(\text{mentions})=0.06$. A seed retained without expansion carries score $0.65 b(s,q)$. Ties are resolved by stable identifier order. The best score across incoming seed edges is used. These constants are fixed in the released experiment script; they are design choices, not optimized test-set estimates.

The last term $a(d,q)$ is an optional sparse topic gate. For each of 20 topic dictionaries, the gate measures normalized overlap between the query tokens and the dictionary of issue descriptions and labels. The top two topic experts are activated, and candidate d receives the sum of their weights when its documented topic membership matches. This is computational sparse routing over handcrafted topical vocabularies; no learned router weights, Transformer experts, or token-level attention is implied. Candidate scoring preserves the original one-hop graph traversal, while making the types of expanded relationships explicit.

To avoid spending the entire context budget on loosely related neighbors, the final selection first reserves the highest-ranked lexical seed and then adds the remaining scored candidates in descending order until $B=5$. Each candidate selected through an edge retains a machine-readable source–relation–target tuple, and validation checks whether the directed edge and its recorded relation type both match E . A deliberately changed edge label is thus detected as a failed provenance assertion even if the endpoints still match. Such validation only tests consistency against the given graph, not whether the underlying legal assertion is authoritative. It is intentionally more modest than factual faithfulness of an LLM explanation.

E. Deterministic algorithm and cost

For each query: (1) tokenize and score all N documents; (2) select three lexical seeds; (3) retrieve one-hop outgoing neighbors; (4) compute typed candidate scores and the optional sparse routing adjustment; (5) reserve the first seed, rank remaining candidates, and return the top five with edge records. A naive inverted index and Python dictionary adjacency lists suffice for the small collection. The BM25 computation is $O(NL_q)$ for query length L_q , while expansion is $O(\text{sum of degrees of the three seeds})$; routing matches the query against 20 short topic dictionaries once per query, caches the two active topic weights, and then looks up each candidate topic membership. This formulation has no claim of scaling to a production legal corpus.

F. Evidence-packet schema and deterministic checks

The output is intentionally an inspectable data object rather than a free-form answer. Each test record contains a stable case identifier, ranked candidate IDs, the complete gold set used by the evaluation process, and a map of selected candidates to source–relation–target tuples. For example, a companion statute can be associated with the seed that caused its graph expansion. This representation separates three otherwise conflated statements: the provision exists in the registry, a typed edge exists in the given graph, and a proposition about the taxpayer is legally supported. The first two can be checked mechanically; the third requires actual source text and professional legal reasoning. Keeping that boundary explicit prevents an attractive diagram from being mistaken for a justification under tax law.

An executable integrity suite applies nine checks to the benchmark and to the outputs of all six retrieval methods. The checks verify uniqueness and registry membership of

returned identifiers, existence and type consistency of provenance edges, consistency of gold identifiers with the registry, disjoint test and development IDs, deterministic repeatability, and independence of retrieval from gold labels, and rejection of a purposely mislabeled graph edge. In particular, replacing a query record’s gold set with a deliberately invalid key does not change the ranking result, because retrieval accepts only query text and index objects. All nine checks pass, and their outcomes are archived with the code. These checks do not substitute for review of whether the registry accurately represents current federal tax authorities.

A machine-readable packet also permits downstream systems to discriminate kinds of insufficiency. A missing source can be detected if the retrieved candidate list lacks a required evidence class; a missing edge is a data-quality fault in the graph; and a mismatched legal conclusion is an inference failure after reading the retrieved text. The controlled corpus has no independent labels for these later failure classes, so only candidate-set omission and mechanically invalid links are evaluated. This decomposition is useful because one global explanation-quality score would conceal which layer needs to be improved.

4. Experimental Design

A. Synthetic benchmark and partition

The controlled benchmark contains 700 generated case descriptions, 35 per topic, using eight sentence templates, six optional filler clauses, and deterministic modifications. Every query stores an exact gold set of one to five section identifiers determined by the scenario-generation rules. The 200 descriptions indexed 00–09 within each topic are designated development material and the remaining 500 descriptions are the test partition. Development labels are not used to fit a model; all reported tables use the test partition only. The seed is 20261022. Because query templates and the graph ontology share a designer, this is a functional validation benchmark with strong construct validity limits, not an independent collection of practitioner-authored questions.

Test queries comprise 280 single-issue scenarios, 160 two-issue descriptions, and 60 deliberately lexically reduced paraphrases. Two-issue examples combine a base topic with another predefined issue and include corresponding evidence identifiers. All legal synopses, query prompts, edge labels, and relevance sets in this evaluation are researcher-authored or programmatically composed; they are not copied from a public test corpus or adjudicated by tax practitioners. The identifiers refer to real provision names, but the benchmark remains entirely synthetic in its evidentiary content.

B. Comparators, ablations, and measures

Methods are TF-IDF, BM25, unrestricted BM25 plus graph expansion, the proposed reranker without relation weights (“no types”), the proposed reranker without topic gating (“no routing”), and the full reranker. All methods read the same corpus, use the same five-document output budget, and are evaluated against the identical gold identifiers. There is no LLM, dense encoder, or supervised risk-classification baseline in the reported experiment. The graph comparator tests specifically the adopted one-hop expansion mechanism from [25], whereas the two ablations isolate our additional scoring decisions.

Recall@5 equals $|P(q) \cap G(q)|/|G(q)|$, precision@5 equals $|P(q) \cap G(q)|/5$, and Complete@5 records whether all gold identifiers are retrieved. We also compute reciprocal rank for

the first relevant returned document, separate average Python wall-clock retrieval time, and graph-link record validity. Differences in Recall@5 are paired by query; confidence intervals use a percentile bootstrap over the 20 issue groups (5000 resamples), avoiding reliance on treating 500 near-template questions as statistically independent. A query-level sign test is retained as an explicitly exploratory diagnostic.

C. Relationship corruption

To probe dependence on relation metadata, the test corrupts selected outgoing expansion labels by replacing them with “mentions” with independent probabilities 0, 0.15, 0.30, and 0.50. Endpoints and corpus text remain unchanged. The unrestricted graph method is therefore invariant to this perturbation by design, whereas the typed method should degrade if its label information becomes unreliable. All values are produced by seeded runs, not extrapolation. The study does not simulate disputed legal precedent or legislative amendments.

D. Structural negative controls and lexical interference

To distinguish the value of relation semantics from an artifact of topological access, we add three predeclared graph perturbations on a balanced 200-query subset of the existing test partition (ten of the 25 test queries per topic). Over six

seeded repetitions, each trial either deletes a random proportion of directed edges, rewires the target of selected edges to another registered identifier while keeping its relation label, or permutes the complete array of relation labels while preserving their global frequency. The untyped graph comparator should be invariant to pure label permutation; typed retrieval should not. These are destructive counterfactual interventions on the synthetic benchmark, not noisy annotations of authentic law.

As a separate full-test query perturbation, we append a sentence explicitly identifying an unrelated tax topic as extraneous administrative information, without changing the original relevance sets. We report the resulting loss of target-source recall and the rate of changed top-ranked identifiers. Finally, an oracle candidate diagnostic counts gold items reachable from the exact three BM25 seeds and their one-hop neighbors. This is an upper bound for pool selection, not a deployable or independently trained retrieval baseline: only the diagnostic sees the gold set.

5. Results

A. Overall retrieval

Table I. Main Controlled Test Results (500 Synthetic Queries; B = 5)

Method	R@5	P@5	Complete@5	MRR
TF-IDF	0.5992	0.4328	0.244	0.922
BM25	0.5993	0.4364	0.216	0.922
BM25 + graph	0.6927	0.4820	0.412	0.795
No types	0.7947	0.5568	0.578	0.917
No routing	0.8461	0.6064	0.590	0.915
TEB-GraphRAG	0.8540	0.6140	0.590	0.915

Table I reports all six methods. The unrestricted graph expansion improves Recall@5 from 0.5993 (BM25) to 0.6927. The typed evidence-budgeted method raises this to 0.8540, a difference of 0.1613. Relative to BM25 the increase is 0.2547. The 95% topic-cluster bootstrap intervals for the

improvements are [0.1137, 0.2160] against the graph comparator and [0.1776, 0.3289] against BM25. These are uncertainty intervals for the synthetic topic collection rather than inference over tax engagements or legal expert populations.

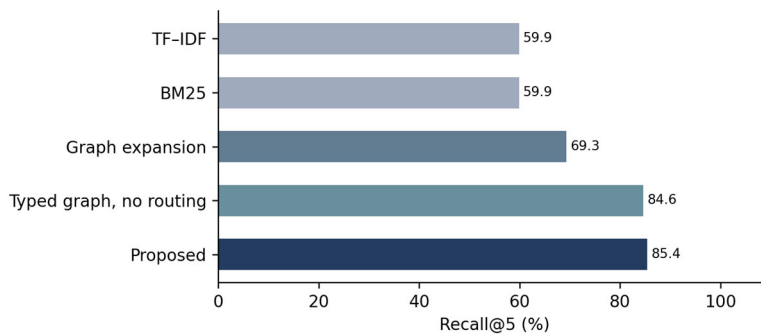


Figure 2. Measured Recall@5 on the 500-case synthetic test set. “No routing” retains type weighting but omits the sparse topic gate.

The complete-evidence rate increases from 0.412 for unrestricted graph expansion to 0.590 for the full method. Reciprocal rank is lower for graph expansion than for lexical retrieval because its top-ranked candidate need not be the most immediately relevant seed; the proposed method restores much of that first-hit behavior with the seed-reservation rule. Because TF-IDF and BM25 have overlapping performance in this small synopsis collection, their difference should not be generalized beyond these data.

B. Ablations and observed mechanisms

Replacing edge-type weights by a constant drops Recall@5 from 0.8540 to 0.7947, an absolute loss of 0.0593. Suppressing topical routing produces 0.8461, a smaller loss of 0.0079. This asymmetry is important: the core gain is attributable to exploiting designed relation categories under the strict retrieval budget, whereas deterministic two-topic gating adds limited incremental value. It would be inappropriate to infer a large neural expert-mixture benefit from this lightweight routing experiment.

The implemented optimization precomputes the top-two

query-topic scores once per query rather than repeating the same dictionary matches for each graph neighbor. Average elapsed Python query times are 0.243 ms for BM25, 0.329 ms for unrestricted expansion, 0.312 ms without routing and 0.521 ms for the full method; caching the topic scores reduced

the latter from 5.279 ms without changing any ranking, an implementation rather than a model-level speedup. These CPU timing figures exclude legal-document download, prompt creation, neural inference, and networking.

C. Robustness to relation-label errors

Table II. Edge-Type Corruption (Recall@5 / Complete@5)

Corrupted labels	Graph expansion	TEB-GraphRAG
0%	0.6927 / 0.412	0.8540 / 0.590
15%	0.6927 / 0.412	0.8233 / 0.526
30%	0.6927 / 0.412	0.7924 / 0.474
50%	0.6927 / 0.412	0.7700 / 0.434

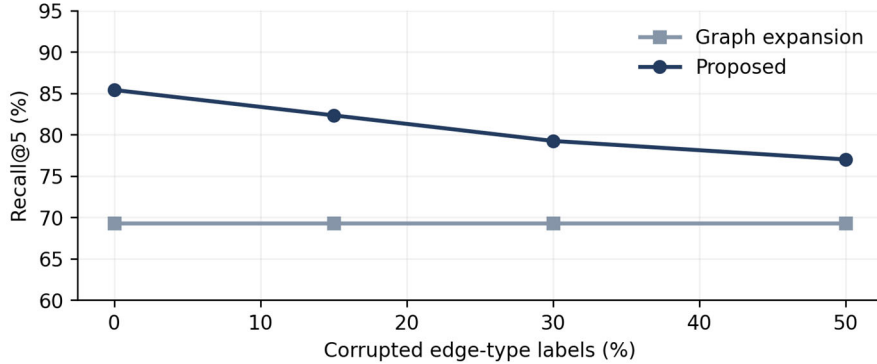


Figure 3. Relation-label corruption leaves untyped expansion unchanged and reduces the advantage of typed expansion.

Under 30% label corruption, Recall@5 drops to 0.7924 and Complete@5 to 0.474, although the method remains above the unrestricted comparator. At 50% corruption the recall is 0.7700. This directional trend is expected: when useful links are falsely relabeled as incidental mentions, their intended weighting advantage disappears. Under the provenance validator, 94.2% of typed packets remain exactly trace-valid

at 30% label corruption; the corresponding untyped expansion packets have 48.8% trace validity because their recorded corrupted labels no longer match the original edge types. This separates ranking robustness from graph-claim integrity. Relation extraction and provenance reliability need independent audits before actual legal deployment.

D. Retrieval-budget sensitivity

Table III. Context-Budget Sensitivity (Recall@B / Complete@B)

Slots B	BM25	Graph expansion	TEB-GraphRAG
3	0.502 / 0.138	0.566 / 0.268	0.739 / 0.580
5	0.599 / 0.216	0.693 / 0.412	0.854 / 0.590
7	0.647 / 0.278	0.783 / 0.560	0.919 / 0.872
9	0.696 / 0.338	0.829 / 0.668	0.930 / 0.910

To test whether the improvement merely reflects an arbitrary choice of five evidence slots, we reran all 500 test queries at budgets B=3, 5, 7, and 9, using the same fixed ranking functions and relevance sets. Table III reports Recall@B and the probability that the retrieved packet contains every gold identifier. The proposed method improves

from 0.739 at B=3 to 0.930 at B=9. The corresponding unrestricted graph method improves from 0.566 to 0.829. Hence the budgeted typed weighting retains an advantage when the evidence allowance is both smaller and larger than the primary setting; this is still an internal controlled comparison rather than a cross-domain result.

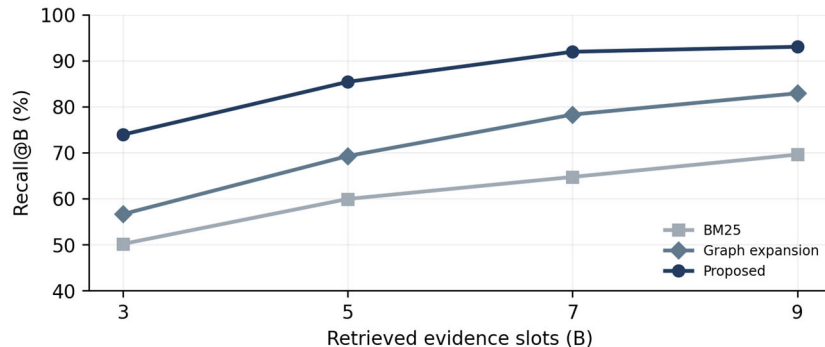


Figure 4. Measured evidence recall as the maximum number of returned authority identifiers increases. All points use the same fixed ranking parameters.

Complete-evidence behavior is especially sensitive to budget: the full method retrieves every gold identifier for 59.0% of cases at B=5, 87.2% at B=7, and 91.0% at B=9. By contrast, graph expansion reaches 41.2%, 56.0%, and 66.8% respectively. The pronounced jump between five and seven slots highlights competition among relevant and irrelevant neighbors. Here two-issue gold sets contain five identifiers and can theoretically fit in B=5, but the observed ranking always omits at least one. Extra slots compensate for this ranking error; the primary bottleneck at B=5 is not gold-set cardinality. Context budgeting should thus be made adaptive to evidence-set complexity rather than fixed without a stopping criterion.

Table IV. Retrieval Recall@5 By Synthetic Case Type

Case type (n)	BM25	Graph	Proposed
Single issue (280)	0.657	0.877	1.000
Two issues (160)	0.600	0.519	0.794
Reduced paraphrase (60)	0.328	0.294	0.333

Table IV exposes a large within-benchmark performance gap. Single-issue cases are recovered completely by the proposed method because the designed typed neighborhood explicitly includes their linked companion provisions. This ceiling is a characteristic of how the synthetic gold sets were constructed, not evidence of perfect real-world retrieval. In two-issue cases Recall@5 is 0.794 and Complete@5 is zero despite each five-identifier gold set being reachable and fitting B=5. This exposes an imperfect joint-evidence ranking rather than an unavoidable output cardinality bound. Reduced paraphrase cases reach only 0.333 Recall@5, little better than BM25, and provide the most persuasive failure mode for a future embedding-based front end.

As a concrete example, query foreign_credit-14 asks about

Table V. Graph Negative Controls (200-Query Balanced Subset; Mean Of Six Runs)

Perturbation	Graph R@5	Typed R@5
Original (unperturbed)	0.713	0.881
Delete 30% of edges	0.650	0.814
Rewire 30% of targets	0.609	0.741
Permute all edge types	0.713	0.597

The unperturbed mean Recall@5 on this 200-query subset is 0.881 for the full method and 0.713 for untyped expansion, distinct from the 500-query main results. Table V summarizes six seeded graph interventions per setting; the complete per-repeat results and repeat standard deviations are preserved. Random deletion of 30% of directed edges reduces typed recall to 0.814; 30% target rewiring reduces it to 0.741. Full relation-label permutation gives 0.597 while the untyped comparator remains 0.713, as expected because the latter ignores type labels. The negative control demonstrates that useful relationship categories, not merely additional connected identifiers, produce the observed ranking advantage on this hand-designed ontology. It also shows that a poorly labeled graph can make typed weighting worse than ignoring types.

Across the 500 original test queries, a sentence marking another tax issue as unrelated nonetheless reduces full-method Recall@5 from 0.854 to 0.636 and changes the first ranked identifier on 53.4% of queries. BM25 similarly falls from 0.599 to 0.474, while untyped graph expansion falls from 0.693 to 0.420. This deliberately adversarial,

These results should not be interpreted as evidence that more context always helps a language model. Recall increases monotonically for nested retrieval lists by definition; retrieved passages may dilute or contradict one another during generation, and token costs rise with the allowance. An operational design would score the marginal information from each additional authority, estimate the chance of a genuinely missing companion, and stop when the extra evidence becomes redundant. This experiment measures candidate-set coverage alone; the question of optimal generative context size is still open.

E. Stratified errors and an auditable example

overseas taxes and cross-border double taxation. Its constructed gold set is IRC-901, IRC-904, and CFR-1.901-2. Unrestricted expansion returns IRC-901 plus four other statutory identifiers, missing both companions. The proposed method returns those three identifiers in its top five, with stored edges IRC-901 → IRC-904 (companion) and IRC-901 → CFR-1.901-2 (interprets). The archived per-query predictions make this check independently repeatable. The words “companion” and “interprets” denote researcher-assigned test labels and must not be interpreted as an authoritative legal characterization of those provisions.

F. Counterfactual controls and candidate-oracle analysis

semantically contradictory phrasing exposes reliance on lexical relevance rather than reliable negation-aware issue selection. It motivates a robust query analyzer and evidence sufficiency checks prior to actual legal use.

The diagnostic oracle finds that the exact three-seed, one-hop candidate pool contains all gold identifiers for every tested single-issue and two-issue query; only the reduced-paraphrase group has a lower mean candidate recall (0.467). Thus the zero two-issue Complete@5 score is predominantly a within-pool ranking and allocation problem, whereas paraphrase failure is partly a seed-generation problem. These are distinct limitations with different remedies: cross-issue set coverage objectives for the former, and nonlexical retrieval for the latter. Oracle values use gold labels strictly after candidate generation and are not executable inference results.

6. Discussion and Implications

The main mechanism is not that graph traversal automatically improves legal understanding. Instead, graph relations act as a scarce-resource allocation signal: five candidate slots should favor explicit designed dependencies

when the initial query is mostly lexical. The improvement over the source-inspired comparator is clearest when a query names one issue but the gold packet also requires related provision identifiers. This is consistent with the purpose of graph-enhanced tax retrieval described in [25], while the controlled evaluation identifies a narrower question—whether distinguishing edge roles matters at fixed output size.

For a production LLM workflow, the packet should be converted into a source-bound context with full authority text, effective dates, exact quoted spans, and a validator that rejects unsupported citations or unsupported propositions. The question to the language model should be whether enough legally effective evidence exists, not simply which retrieved section has the largest score. A human reviewer should be able to inspect the retrieval path, source hierarchy and any conflicting authority. Those components are design implications, not outcomes measured in this paper. Rhee et al. provide a complementary experimental perspective through self-play contrastive fine-tuning of web search agents, where search trajectories furnish structured evidence for training and evaluation [27].

A real empirical extension would first replace synopsis metadata with a versioned Title 26 U.S.C./C.F.R. corpus, extract references from primary texts, and independently review typed edges. It would then evaluate real practitioner-written UTP narratives with expert-defined source judgments, both at the passage level and for source-grounded explanations. In that setting, missing sources and legally obsolete links may be more costly than topical false positives; cost-sensitive relevance and abstention may be appropriate. Comparing an actual domain-tuned legal LLM to a frozen LLM supplied with verified evidence would be a separate, meaningful generation study.

A. Relation governance and source hierarchy

Legal relation labels must be interpreted more carefully than ordinary semantic links. A regulation’s relationship to a statutory section is not equivalent to the relationship between two statutes that discuss neighboring issues, and neither implies that a tax position is sustainable on its specific facts. Our synthetic types are useful ranking features because their class frequencies and strengths are known by construction. They are not a substitute for a hierarchy of legislative authority, judicial precedential weight, regulatory effective dates, and administrative guidance. A realistic system would encode the citation origin, relation extractor version, confidence, effective period and human-review status for every edge. The scoring weight could then be a calibrated estimate of source usefulness conditional on that metadata rather than an unqualified fixed constant.

The label-corruption study approximates the loss of a useful relation class; our additional edge-deletion and endpoint-rewiring controls test two structural failure modes. None approximates time-expired statutory provisions, conflicts among genuine authorities, or jurisdictional precedential hierarchy. These failures could interact with the budget in nontrivial ways. For instance, a missing cross-reference may prevent candidate generation altogether, whereas an outdated but high-scoring link may actively push current guidance out of the retrieved packet. A future deployment evaluation should therefore include both edge-label noise and structural graph errors, anchored to a legally versioned ground truth.

B. When to abstain rather than retrieve more

High recall is useful only if the selected sources answer the

right question. A system can collect every section in a designer-authored gold set yet fail to represent a factual exception, a jurisdictional limitation, or an unfavorable authority. In a compliance workflow, the retriever should flag low-confidence matches, incompatible as-of dates, and missing documentation instead of forcing a risk label. A language model receiving the evidence packet could be constrained to respond with a request for additional facts whenever a dispositive requirement is absent. Evaluating this behavior requires tax-professional judgments on both the relevance and the sufficiency of evidence, which are not furnished by the present synthetic labels.

The observed results suggest a staged research sequence. First, audit graph construction and reference resolution on real federal law. Second, evaluate evidence coverage under document-level and passage-level judgments made independently of graph creation. Third, add a generative model with strict source quoting and proposition-level citation checks, measuring factual consistency, refusal appropriateness and human usefulness. Only after these stages would ordinal risk assessment or uncertain-tax-position decision support warrant empirical claims. This sequence deliberately avoids conflating component reliability with end-to-end legal correctness.

7. Validity and Reproducibility

The benchmark is small and intentionally structured: 59 section identifiers, a hand-designed graph, repeated template families, and programmatically determined relevance. It tests the correctness of an algorithmic transformation under graph assumptions, not its ecological validity. A graph designed around gold dependencies favors relation-aware methods, making the single-issue perfect score especially prone to optimism. The two-issue ranking failure, lexical distractor sensitivity, type-permutation collapse, and weak paraphrase results show where this constructed advantage does not resolve harder retrieval problems. Repeated perturbation trials quantify synthetic-graph sensitivity, not confidence intervals for any legal population.

The experiment includes neither raw government legal text nor independent expert assessment. Section names are real, but synopses are simplified researcher-authored descriptions and relation labels are manually posited. Their semantics are not legally certified. No generated explanation was reviewed, no uncertainty-position classification was trained, no LLM was invoked, and no predictive accuracy is asserted. Mechanical path validity does not measure legal entailment, current authority, jurisdictional precedence, or citation precision in generative output.

Reproducibility artifacts include the complete script, seed, fixed scoring coefficients, authority registry, query partition, typed graph edges, per-query predictions, summary metrics, budget sensitivity, confidence intervals, corruption results, seeded topology and label counterfactuals, lexical distractor pairs, and candidate-oracle diagnostics. Three scripts reconstruct the metrics, the budget analysis, and the additional controls offline. Dependencies are NumPy, SciPy and scikit-learn. Query labels are held separately from ranking inputs in each call; the ranking code receives only query text, the synopsis registry, the graph and the fixed parameters. An independent legal-data study remains necessary before any practical tax-risk assessment claim.

8. Conclusion

This work develops and evaluates a typed, evidence-budgeted tax graph retriever grounded in a concrete graph expansion mechanism. On 500 controlled synthetic scenarios, it improves Recall@5 from 0.6927 for unrestricted graph expansion to 0.8540, with large degradation after label permutation and exposure to unrelated query content. Two-issue evidence ranking and paraphrase seed formation remain unresolved. The contribution is a transparent and testable retrieval-stage modification, not a validated tax-risk classifier or an evaluated LLM. Subsequent research should prioritize authentic source documents, expert relevance judgments, legal temporal validity, and generation grounded in auditable evidence paths.

References

- [1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Galkin, A., Goyal, N., Ott, M., Grave, E., Lample, G., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://arxiv.org/abs/2005.11401>.
- [2] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1–2), 1–174. <https://doi.org/10.1561/15000000019>
- [3] Edge, D., Truong, H., Miller, B., & Gokhale, A. (2024). From local to global: A Graph RAG approach to query-focused summarization. <https://arxiv.org/abs/2404.16130>
- [4] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. *World Journal of Management Science*, 2(1), 63–72.
- [5] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The Muppets straight out of Law School. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2898–2904. <https://doi.org/10.18653/v1/2020.findings-emnlp.261>
- [6] Chalkidis, I., Manginas, A., Katakis, E., Kotsis, G., Aletras, N., & Androutsopoulos, I. (2022). LexGLUE: A benchmark dataset for legal language understanding in English. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4310–4330. <https://doi.org/10.18653/v1/2022.acl-long.297>
- [7] Guha, N., Block, S., Wang, J., Martin, M., Hirst, G., Ho, D., Grisham, C., & Mitchell, T. (2023). LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36. <https://doi.org/10.52202/075280-1915>
- [8] Shao, Y., Mao, J., Liu, Y., He, Y., Yan, R., Zhang, M., & Ma, S. (2020). BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 3501–3507. <https://doi.org/10.24963/ijcai.2020/484>
- [9] Karpukhin, V., Oguz, B., Min, S., Lyu, X., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [10] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992. <https://doi.org/10.18653/v1/D19-1410>
- [11] Zhu, X., Xie, Y., Liu, Y., Li, Y., & Hu, W. (2025). Knowledge graph-guided retrieval augmented generation. *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8912–8924. <https://doi.org/10.18653/v1/2025.naacl-long.449>
- [12] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. *Innovation and Technology Studies*, 3(1), 27–41.
- [13] Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., & Manning, C. D. (2024). RAPTOR: Recursive abstractive processing for tree-organized retrieval. *Proceedings of the International Conference on Learning Representations*.
- [14] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of the International Conference on Learning Representations*.
- [15] Wang, J., Ge, Y., Liu, J., & Jiang, B. (2024). Early warning of liquidity stress in U.S. banks with seasonally aligned frailty hazards and delayed-feedback conformal alarms. *Journal of Trends in Finance and Economics*, 1(1), 85–100.
- [16] Gao, L., Schulman, J., Hilton, J., Madaan, A., Kim, Y., Rao, N., Szlam, A., & Wu, J. (2023). RARR: Researching and revising what language models say, using language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16477–16508. <https://doi.org/10.18653/v1/2023.acl-long.910>
- [17] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100. <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- [18] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *Proceedings of the International Conference on Learning Representations*.
- [19] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. <https://arxiv.org/abs/1701.06538>
- [20] Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120), 1–39.
- [21] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W., & Chen, Z. (2022). LoRA: Low-rank adaptation of large language models. *Proceedings of the International Conference on Learning Representations*. <https://arxiv.org/abs/2106.09685>
- [22] Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. <https://doi.org/10.1109/TBDATA.2019.2921572>
- [23] Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758–759. <https://doi.org/10.1145/1571941.1572114>
- [24] Teng, D. (2025). TRIAD: Deadline- and reliability-aware scheduling for post-earthquake structural assessment workflows. *World Journal of Engineering Research*, 3(6), 55–66.

- [25] Tan, Y., Wu, B., Cao, J., & Jiang, B. (2025). LLaMA-UTP: Knowledge-guided expert mixture for analyzing uncertain tax positions. *IEEE Access*, 13, 90637–90650. <https://doi.org/10.1109/ACCESS.2025.3571502>
- [26] Zhao, W., Chen, T., Wang, B., & Shen, Z. (2026). PIPE-RAG: An execution-guided retrieval and reuse framework for reliable machine learning pipeline code generation. *WEP Transactions on Computer Science and Intelligent Systems Research*, 13, 286–297.
- [27] Rhee, M., Kim, J., Park, S., & Lee, J. (2026). Enhancing Web search agents with self-play contrastive fine-tuning. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2026.3717594>