

Citation Graph-Based Multi-Hop Retrieval Optimization Using Hybrid RRF in a RAG System for Indonesian ITE Law Documents

Andi Laksono^{1*}, Nova Rijati^{2*}

* Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang, Indonesia
andilaksono70@gmail.com¹, nova.rijati@dsn.dinus.ac.id²

Article Info

Article history:

Received 2026-09-11

Revised 2026-10-14

Accepted 2026-10-17

Keyword:

*Retrieval-Augmented
Generation,
Citation Graph Expansion,
Hybrid Retrieval,
Legal Documents,
UU ITE,
Multi-Hop Retrieval.*

ABSTRACT

Indonesian legislation, including the Electronic Information and Transactions Law (UU ITE), separates prohibition articles from sanction articles into distinct sections while linking them through explicit cross-references. This structural pattern creates multi-hop retrieval challenges that conventional Retrieval-Augmented Generation (RAG) systems relying on semantic and lexical similarity alone fail to resolve. This study proposes and evaluates a RAG system for UU ITE legal documents combining hybrid retrieval based on Reciprocal Rank Fusion (RRF) with citation-graph expansion across articles. Evaluation was conducted on a consolidated corpus of 63 articles spanning the 2008, 2016, and 2024 regulatory layers, using 40 manually validated query-ground truth pairs across three conditions: a hybrid RRF baseline, hybrid RRF with citation-graph expansion, and a quantity-matched RRF control isolating graph selectivity from mere chunk-quantity effects. Results show that citation-graph expansion raised overall recall to 0.948 and Full-Hit Rate to 0.875, higher than both the baseline (0.904 and 0.800) and the quantity-control condition (0.910 and 0.800), despite the latter retrieving 34% more chunks. Stratified analysis confirms the gains are concentrated in multi-hop queries with explicit graph connections (Full-Hit Rate rising from 0.500 to 0.714), with no effect on single-hop or edgeless multi-hop queries. A Wilcoxon signed-rank test on this stratum did not reach statistical significance ($r = 1.000$; $p = 0.059$) due to the small sample size; the results are therefore suggestive rather than conclusive, and larger-scale replication is needed before firmly establishing the effectiveness of citation-graph expansion.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

The Indonesian Electronic Information and Transactions Law (UU ITE), like statutory legislation in general, is drafted in normative and formal language, with lengthy article structures and an extensive network of cross-references. These characteristics often prevent laypeople from recognising the boundaries of conduct regulated by the law, so an act that seems ordinary in everyday digital interaction may lead to unanticipated legal consequences. The growth of digital information has driven demand for systems that can manage and present information quickly, in a structured and accessible manner. Question-answering systems based on Retrieval Augmented Generation (RAG) have emerged as one approach with the potential to bridge this comprehension gap by integrating a document retrieval mechanism with the

generative capability of large language models (LLMs), so that the generated answers can be traced back to the original legal text rather than relying solely on the model's internal parameters, which are prone to producing unsubstantiated information [1]. However, applying generic RAG to legal documents presents distinct challenges, given that legal text has a strictly formal structure and a network of references that is not always captured by retrieval mechanisms based on semantic similarity alone [2].

One of the fundamental challenges in applying RAG to legal documents lies in the chunking strategy employed. Study [1] shows that generic, fixed character length chunking methods outperform third-party reranking approaches in the legal domain, indicating that mechanisms designed for general-purpose domains are not necessarily well suited to the structural characteristics of legal text. This issue is

compounded by the phenomenon of Document-Level Retrieval Mismatch (DRM), a discrepancy between the document segments successfully retrieved by the system and the document context actually required, as identified by [3] in long-form legal documents. Beyond these structural issues, retrieval based purely on semantic similarity is also prone to failing on queries that use legal terms with different morphological forms but equivalent meaning, a problem that is more pronounced in agglutinative languages such as Indonesian. Both issues indicate that retrieval accuracy in the legal domain cannot be resolved simply by scaling up the model or the training data, but instead requires adjustments to the preprocessing strategy and the document retrieval stage itself.

In addition to the challenges at the chunking and basic retrieval stages, codified legal documents exhibit a relational characteristic that cannot be handled by text- or vector similarity-based retrieval alone: an explicit network of cross-references between articles. Study [4] shows that this complexity increases further in regulations that have undergone amendment over time, where conventional retrieval systems risk retrieving a version of the text that is no longer in force, absent a mechanism for tracking revision history. Efforts to address cross-referencing through knowledge-graph fusion have shown considerable promise. Study [5] reports an F1-score improvement of up to 111% over a BM25 baseline when hybrid retrieval combining BM25, dense semantic search, and graph search was applied to trace cross-citations across 2.47 million Swiss court documents. However, this substantial improvement was validated on an extremely small test query set (only 10 queries) and required heavy, iterative GPU-based computation, such that the benefits of citation-graph mechanisms at that scale cannot necessarily be replicated efficiently on more limited infrastructure or data scale. This underscores that purely semantic or lexical retrieval is indeed inadequate for capturing complex legal relations, yet existing graph-based solutions still leave open questions regarding validity at small evaluation scales and computational efficiency under limited resources.

Although several knowledge-graph-based approaches have been proposed to address the cross-reference problem, a fundamental weakness in the existing literature lies in the lack of staged validation of each pipeline component's contribution in isolation. In the Indonesian context specifically, a study [6] evaluated hybrid BM25 and dense retrieval via Reciprocal Rank Fusion on 104 documents from Surabaya's Regional Regulations (Peraturan Daerah), but obtained comparatively low performance ($F1 = 0.2867$) and explicitly restricted their testing to single-hop queries with a static chunking size, a limitation they recommend addressing through multi-hop query testing and more adaptive chunking strategies. A similar effort in the Indonesian criminal law domain by [7] reported promising results (an 80% hit rate), but the corpus used was limited to a single draft piece of legislation, and therefore did not test how the system handles the inter-article

cross-reference networks that characterise codified regulations such as UU ITE. The absence of a staged ablation study separating the contributions of chunking, hybrid retrieval, and the citation-graph mechanism makes it difficult to determine which component is genuinely responsible for the performance gains reported in prior studies. This gap is made more pressing by the fact that multi-hop legal queries are precisely the limitation identified by [6] are a common scenario in UU ITE, where the criminal provisions in one article are structurally dependent on definitions or prohibited acts set out in another article.

While prior RAG research has examined hybrid and graph-based retrieval independently, this study specifically applies a citation graph to resolve a structural challenge inherent to codified legislation, namely the systematic separation of prohibition provisions from their corresponding sanction provisions linked via explicit cross-references. Unlike general-purpose graph-RAG approaches [5] that apply broad graph fusion over heterogeneous court documents without attributing performance gains to specific pipeline stages, and existing Indonesian legal RAG studies [6], [7] that omit multi-hop cross-reference evaluation, this study introduces a rigorous staged ablation framework. By comparing baseline RRF, graph-expanded RRF, and a quantity-matched RRF control, this framework isolates true citation-graph selectivity from mere retrieval-depth effects, addressing a key methodological gap in prior work [5], [6], and [7].

Building on these gaps, this study proposes a RAG pipeline for UU ITE that combines per-article chunking, hybrid retrieval (BM25 with the Sastrawi stemmer and FAISS, combined via Reciprocal Rank Fusion), and a post-retrieval citation-graph expansion mechanism to handle cross-references between articles. Unlike large-scale citation-graph approaches that demand high computational resources, this study is designed to be validated at a more measured evaluation scale by explicitly distinguishing performance on single-hop queries from that on multi-hop queries requiring cross-reference traversal. This distinction has not been previously examined in prior Indonesian language legal RAG research. The contribution of each pipeline component is tested in a staged manner through an ablation study, rather than being assumed to contribute positively by default, thereby making it possible to determine which component is genuinely responsible for the observed performance gains. This approach is intended to provide more decisive empirical evidence regarding the feasibility of citation-graph mechanisms in the Indonesian codified legal domain, while also offering a replicable evaluation framework for other Indonesian regulations.

II. METHODS

This section describes the methodology used in this research. The research process consists of several stages; the overall workflow of the proposed methodology is presented in Figure 1 below.

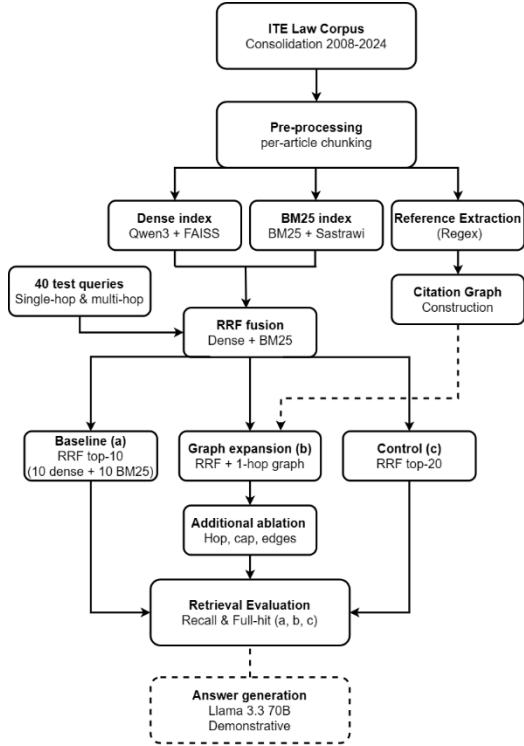


Figure 1. Research Flow

A. Corpus and Preprocessing

The corpus for this study is drawn from an unofficial consolidated text of the Electronic Information and Transactions Law (UU ITE), which merges Law No. 11 of 2008, Law No. 19 of 2016 (first amendment), and Law No. 1 of 2024 (second amendment) into a single document. The text was obtained through digitisation (OCR) with manual correction of clear typographical errors. This study did not perform systematic cross-verification against the official State Gazette (Lembaran Negara), so the accuracy of the text depends entirely on the quality of the consolidated source used.

Preprocessing was carried out by splitting the document into text units (chunks) at the article (pasal) level, using pattern-based segmentation on [PASAL N]. Each chunk was annotated with a legal status (original, amended, or inserted) to track amendment history, and filtered to include only sections marked [JENIS: NORMA], thereby excluding the Penjelasan (non-normative interpretive commentary) from the retrieval corpus. This process yielded 63 unique chunks.

Because the corpus is limited to a single regulatory domain (UU ITE) comprising 63 articles, the findings reported in this study should be interpreted as a proof-of-concept demonstration of citation-graph-based multi-hop retrieval on codified Indonesian legislation, rather than as generalizable evidence across other regulatory domains, larger corpora, or statutes with denser or more heterogeneous cross-reference structures. Replication on additional

Indonesian laws is necessary before broader claims of applicability can be made.

The per-article chunking strategy was chosen based on two considerations. First, several articles contain clauses (ayat) that are structurally interdependent, so splitting at the clause level risks producing chunks that lose their grammatical antecedent [8]. Second, the explicit cross-referencing mechanism between provisions in Indonesian legislative text operates at the article level, so this granularity aligns directly with the unit used by the citation-graph expansion mechanism [9].

B. Dense Retrieval

The dense retrieval component was built using the Qwen-0.6B embedding model. This model was selected based on three practical considerations. First, a model with 0.6 billion parameters can be run entirely on the NVIDIA T4 GPU (14.56 GB VRAM) available in the Google Colaboratory environment without memory allocation failures. Second, the model is designed to support multilingual text, including Indonesian, and is therefore expected to capture the semantic nuances of Indonesian legal terminology better than embedding models built exclusively for English [10]. Third, the model's compact size yields low inference cost, which is relevant for the repeated-use scenario required by systematic evaluation.

$$\hat{v} = \frac{v}{|v|_2} \quad (1)$$

Before indexing, each representation vector v for both documents and queries is normalised using L2 normalisation as in Equation (1), where $|v|_2$ denotes the Euclidean norm of vector v [11]. This normalisation produces a unit-length vector $\hat{v}$ such that the inner-product computation at the search stage becomes mathematically equivalent to cosine similarity [12].

$$\text{sim}(q, d) = \hat{v}_q \cdot \hat{v}_d = \cos(\theta) \quad (2)$$

The similarity score between query q and document d is computed as the inner product of the normalised query vector $\hat{v}_q$ and the normalised document vector $\hat{v}_d$, as in Equation (2). Because both vectors have been normalised to unit length, this inner-product value is mathematically equivalent to the cosine of the angle θ between them, such that using FAISS (an inner-product-based index) yields a retrieval ranking identical to that of a conventional cosine-similarity approach, without requiring additional norm computation at search time [12], [13].

For indexing and search, this study uses the IndexFlatIP index structure from the FAISS (Facebook AI Similarity Search) library. This index type performs exact search by computing the inner product against every vector in the corpus exhaustively [13]. This index was selected given the very small corpus size of 63 article chunks.

C. BM25 Retrieval

The sparse retrieval component was implemented using the BM25 algorithm, specifically the BM25Okapi variant [14]. Default parameter values were used, without hyperparameter tuning.

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + K} \quad (3)$$

$$K = k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgl}} \right) \quad (4)$$

This computes the relevance score of document d with respect to query q by summing the contribution of each query term. Each term is weighted according to its inverse document frequency (IDF), which suppresses the influence of common words appearing across many articles while amplifying the influence of rarer, topically specific words. The component K in Equation (4) combines the term-frequency saturation parameter k_1 with document-length normalisation via parameter b , so that documents substantially longer than the corpus average do not automatically receive a higher score merely by containing more repeated words. This study uses $k_1 = 1,5$ dan $b = 0,75$, the library's default values, without additional tuning to the characteristics of the UU ITE corpus [14].

Before indexing with BM25, each article chunk passes through a three-stage tokenisation pipeline. First, all text is converted to lowercase [15], [16]. Second, non-alphanumeric characters are removed using regular expressions, leaving only letters, digits, and spaces. Third, each token is stemmed using PySastrawi, a morphological stemmer for Indonesian that reduces words to their root form [17]. The same pipeline is applied to query text at search time, so that query and document representations remain consistent within the same lexical space. BM25 retrieves the top 10 candidates ranked by relevance score, which are subsequently combined with the dense retrieval results through the RRF mechanism in the following stage.

D. Hybrid Retrieval Based on Reciprocal Rank Fusion

The outputs of dense retrieval and BM25 are combined using Reciprocal Rank Fusion (RRF). RRF serves as the base fusion mechanism used across all ablation conditions in this study. The conditions differ not in the fusion formula itself, but in the top-k output parameter and in whether citation-graph expansion is applied after fusion. RRF was chosen because it does not require score normalisation across retrieval systems operating on different scales, since directly summing scores from systems with incompatible scales is not valid without additional calibration [18], [19].

The RRF score for each document is computed as:

$$\text{RRF_Score}(d \in D) = \sum_{m \in M} \frac{1}{k + r_m(d)} \quad (5)$$

where M is the set of retrieval systems (dense and BM25), $r_m(d)$ is the rank of document d in system m , and k is a smoothing constant [18]. This study uses $k = 60$, following the default value reported in the original literature and shown to be robust across a range of document collections.

The implementation takes the top-k candidates from each system, computes the RRF score for every document appearing in either or both lists, then returns the documents ranked by their RRF score in descending order. Documents appearing in both lists naturally receive a higher score than documents appearing in only one system.

E. Citation Graph Construction

A citation graph was constructed to represent explicit reference relationships between articles in the consolidated UU ITE corpus. The data structure used is a pair of native Python dictionaries, since the small corpus scale does not warrant the overhead of a full graph library [20].

Reference extraction was performed via regular expressions applied to each chunk's text [20]. This construction yielded 49 directed edges from 26 chunks with at least one outgoing reference. When considered as an undirected graph, there are 47 mutually connected article pairs. The remaining 37 chunks have no outgoing references and therefore do not contribute to forward expansion; however, they remain eligible targets for reverse expansion if referenced by other articles [21]. Because the citation graph is constructed over the single consolidated corpus described in Section Corpus and Preprocessing in which each article retains a single final numbering regardless of its amendment history, regular expression reference extraction inherently resolves against current in-force article numbers, thereby avoiding stale or ambiguous edges to superseded article versions.

Graph expansion at retrieval time proceeds in two directions. Forward expansion traces the articles referenced by a successfully retrieved chunk [20], [21]. Conversely, reverse expansion retrieves other articles that explicitly reference the target chunk. This design specifically addresses a structural characteristic of Indonesian legislation, wherein prohibition articles and their corresponding sanction articles are positioned far apart within the text yet remain connected through explicit references [22], [23], [24]. Without reverse expansion, standard retrieval systems fail to answer queries requiring both provisions, as conventional vector search cannot bridge textually distant yet structurally dependent clauses. As demonstrated empirically by the ablation results in Section Additional Ablation, reverse expansion alone accounts for the highest system performance, whereas forward expansion alone provides no measurable gain over the baseline without graph expansion.

F. Experimental Design

To evaluate the contribution of citation-graph expansion in a controlled manner, this study designed three retrieval conditions compared within an ablation framework [19]. All three conditions share the same RRF fusion mechanism; they differ only in the top-k output parameter and in whether graph expansion is applied after fusion [18].

1) Baseline Hybrid RRF (Condition a)

The baseline condition combines dense retrieval and BM25 via RRF, producing 10 ranked chunks that are passed directly to the generation model without further modification. This condition represents a standard hybrid RAG system without the proposed component.

2) Baseline with Citation Graph Expansion (Condition b)

The treatment condition begins from results identical to condition (a), then applies single-hop citation-graph expansion with a maximum of five additional articles per query. Each chunk obtained through expansion is assigned a placeholder score of 0.5, since all expansion occurs at the first hop. Expanded chunks are appended to the end of the list after the RRF-ranked chunks, without a combined re-ranking mechanism, so the priority ordering of the RRF chunks is preserved. The realised average number of chunks passed to the LLM under this condition is 14.95 per query, higher than condition (a)'s fixed 10, since expansion adds chunks beyond that limit.

3) Quantity Control (Condition c)

The control condition is specifically designed to isolate the effect of chunk quantity from the effect of graph-based structural selectivity. This condition runs pure RRF without graph expansion. Dense retrieval takes 20 candidates, BM25 takes 20 candidates, and the RRF output is capped at 20 chunks. No other modification is applied.

Both conditions (b) and (c) derive their candidate pools from identical top-20 BM25 and top-20 dense retrieval rankings prior to fusion, without introducing additional retrieval passes. They differ exclusively in how supplementary chunks are selected. Condition (b) selects up to five chunks connected via explicit citation edges, whereas condition (c) retrieves up to 20 chunks by extending further down the existing RRF ranking. Consequently, both conditions are controlled for retrieval depth and candidate pool origin, isolating graph connectivity from RRF rank as the key selection criterion for extra chunks.

The realised average chunk count under this condition is 20, quantitatively larger than condition (b)'s 14.95. This design is conservative with respect to the claims made: if condition (b), with only 14.95 chunks, still yields higher recall than condition (c) with 20 chunks, then the performance improvement cannot be explained merely by the quantity of context provided to the LLM [25], [26]. Any remaining difference can then be attributed only to the relevance of the additional chunks selected through the citation-graph structure.

G. Answer Generation

The generation stage uses the Llama 3.3 70B Versatile model, accessed via the Groq API. The model was selected for its fast, free-tier inference availability on Groq [27], as well as its sufficient context capacity to accept up to 20 article chunks at once [28].

The context passed to the model is assembled from the retrieved chunks with source-label formatting. Chunks originating from citation-graph expansion are given an additional marker to distinguish them from chunks obtained directly through RRF. This marker provides the model with transparency regarding the provenance of each context segment, while also enabling manual auditing of each component's contribution to the final generated answer [29].

The generation stage in this study serves as an end-to-end demonstration to verify that the pipeline is functionally operational and that the retrieved chunks can indeed produce coherent, legally relevant answers. Formal evaluation of answer quality is outside the scope of this study and is identified as a direction for future work.

H. Evaluation Data and Metrics

The evaluation data consists of 40 query-ground-truth pairs generated using the Gemini 3.5 Flash language model as an initial drafting aid. Following automated generation, all 40 queries were manually validated by the researcher by matching each cited article against the actual article text in the consolidated file. The validation process resulted in six corrections to incorrect article numbers and one query and ground truth replacement, for a total of seven items revised from the LLM's initial draft [30]. Validation was performed by a single researcher without a second, independent validator, so the possibility of annotation bias cannot be quantified through inter-annotator agreement.

The data is divided into two groups based on the retrieval characteristics required. Twenty queries are single-hop, where the answer can be found within a single article without requiring cross-referencing. Twenty queries are multi-hop, where the answer requires synthesising information from two or more mutually referencing articles [30], [31]. To sharpen the analysis, multi-hop queries were further grouped according to the presence of an explicit edge in the citation graph. Of the 20 multi-hop queries, 14 have at least one ground-truth article pair explicitly connected via the citation graph, and are therefore positioned to benefit directly from the graph expansion mechanism. The remaining six queries have no explicit graph connection; retrieval for these queries depends entirely on the system's semantic and lexical capability, without the aid of graph expansion [32].

TABLE I
QUERY CATEGORY BREAKDOWN

Category	n	Percentage
Single-hop	20	50%
Multi-hop+edge	14	35%
Multi-hop-no-edge	6	15%
Total	40	100%

A retrieved set was defined as a complete legal answer if it met five specific criteria: first, it contained the prohibition or normative article directly addressing the query; second, when the query involved legal consequences, it included the corresponding sanction article targeted by citation-graph expansion; third, it incorporated a definitional article only when the defined term was critical to resolving the query, excluding articles containing merely generic definitions; fourth, for articles amended or modified by the 2016 or 2024 laws, only the current in-force version consistent with the chunk metadata was accepted as ground truth; and fifth, articles that were merely thematically related without explicit cross-references were excluded as required items. This exclusion rule serves as the operational criterion distinguishing multi-hop queries with edges from those without edges, rather than functioning as a post-hoc label.

Evaluation was performed at the retrieval level, prior to the generation stage. Two metrics were used: Recall@K measures the proportion of ground-truth articles successfully included in the list of chunks passed to the LLM.

$$\text{Recall} = \frac{|P_{\text{expected}} \cap P_{\text{retrieved}}|}{|P_{\text{expected}}|} \quad (6)$$

Recall is computed as the proportion of ground-truth articles (P_{expected}) found among the retrieved articles ($P_{\text{retrieved}}$), as in Equation (6). This metric is commonly used to evaluate the retrieval component of a RAG system independently of the quality of the LLM's final answer, as it directly measures the system's ability to locate relevant source documents without being influenced by the language model's generative capability [33].

$$\text{FullHit} = \mathbb{1}[\text{Recall} = 1] \quad (7)$$

Full-Hit Rate is a binary per-query indicator, equal to 1 if all ground-truth articles are successfully retrieved (Recall = 1) and 0 otherwise, as in Equation (7). This metric is particularly important for multi-hop queries that require more than one article simultaneously. Partial recall (e.g., finding 1 of 2 required articles) is insufficient to answer the question fully, so Full-Hit Rate provides a stricter picture of the system's ability to answer cross-article queries completely [33], [34].

$$W = \min(W^+, W^-) \quad (8)$$

The significance of differences between conditions was tested using the Wilcoxon signed-rank test, a non-parametric test for paired data suitable for small sample sizes and free of normal-distribution assumptions. The test statistic W is computed as the minimum of the sum of positive ranks (W^+) and the sum of negative ranks (W^-) of the paired differences between the two conditions being compared, as in Equation (8). The

resulting p value was used to determine significance at the threshold $\alpha = 0,05$ [35], [36].

$$r_{\text{tb}} = 1 - \frac{2W}{n_{\text{eff}}(n_{\text{eff}} + 1)} \quad (9)$$

To quantify effect size beyond statistical significance alone, rank-biserial correlation (r_{tb}) is also reported, computed from the statistic W and the number of non-tied pairs (n_{eff}), as in the equation above. The value of r_{tb} ranges from 0 to 1, with values approaching 1 indicating a consistently strong effect favouring one condition. n_{eff} is reported explicitly because a high tie ratio relative to the nominal sample size has a direct bearing on the statistical power available [37].

Testing was run separately across three strata: single-hop, multi-hop+edge, and multi-hop-no-edge. Two comparisons were tested: (a) vs. (b), to measure the contribution of graph expansion, and (b) vs. (c), to isolate the effect of graph selectivity from the effect of chunk quantity only.

III. RESULT AND DISCUSSION

A. Corpus and Citation Graph Statistics

The corpus used in this study consists of 63 article (pasal) chunks extracted from the consolidated UU ITE text file, which spans three regulatory layers: Law No. 11 of 2008 as the principal body of the law, Law No. 19 of 2016 as the first amendment, and Law No. 1 of 2024 as the second amendment. Each chunk represents one complete article. Of the 63 articles, 44 (69.8%) remained unchanged from the 2008 version, 11 (17.5%) underwent substantive revision under the 2016 or 2024 amendments, and 8 (12.7%) are new articles inserted by the 2024 amendment, namely Articles 13A, 16A, 16B, 18A, 27A, 27B, 40A, and 45B. This composition reflects the fact that the 2024 amendment substantively expanded the law's scope, particularly with respect to child protection in digital spaces, electronic extortion and threats, and electronic certification.

The citation graph was constructed from the explicit inter-article references found in the corpus text. This construction yielded 49 directed edges from 26 chunks with at least one outgoing reference, meaning 41.3% of all chunks contribute to the graph structure. These references reach 27 unique target articles, indicating that more than half of the articles in the corpus participate in the reference network, either as source or as target.

Article 27 is the most frequently referenced article, with 9 incoming references. Consistent with its position as the principal prohibition article, which serves as the reference point for several sanction and aggravation articles. The next four most referenced articles are 27A, 28, 30, and 36, each referenced 3 times. This concentration of references around prohibition articles reflects a structural characteristic typical of Indonesian criminal legislation, in which sanction articles explicitly reference back to the prohibition article on which they are based.

B. Retrieval Evaluation Results

Table II presents the evaluation results for the three retrieval conditions across all 40 queries. Condition (b), which combines hybrid RRF with citation-graph expansion, achieves the highest recall (0.948) and Full-Hit Rate (0.875), surpassing both baseline condition (a) and quantity-control condition (c).

TABLE II
RETRIEVAL EVALUATION RESULTS

Condition	Recall@K	Full-Hit Rate	Average Chunks
BM25-only	0.827	0.725	10.00
Dense-only	0.879	0.775	10.00
(a) Baseline RRF top-10	0.904	0.800	10.00
(b) RRF + Citation Graph	0.948	0.875	14.95
(c) RRF top-20 control	0.910	0.800	20.00
Upper bound all chunks	1.000	1.000	63.00

Before discussing the contribution of graph expansion, it is first necessary to establish the empirical justification for adopting hybrid retrieval as the system's foundation. A comparison between the single-retriever components shows that dense retrieval slightly outperforms BM25 (a difference of +0.052 in recall and +0.050 in Full-Hit Rate), indicating that semantic embedding matching is more effective than purely lexical matching for this legal-text domain, since phrasing variation between queries and article wording is not always captured by keyword overlap. However, the hybrid condition (a), which combines both via RRF, surpasses both single components (recall of 0.904 versus 0.879 for dense-only and 0.827 for BM25-only), providing empirical rather than merely theoretical justification that combining lexical and semantic signals is complementary for legal document retrieval.

The next finding concerns the comparison between condition (b) and condition (c). Condition (c) retrieves an average of 20.00 chunks per query, 34% more than condition (b)'s 14.95 chunks, yet yields recall and Full-Hit Rate identical to baseline (a). In other words, doubling the number of chunks passed to the LLM through pure RRF provides no meaningful performance gain. Condition (b), by contrast, achieves a recall improvement of +0.044 and a Full-Hit Rate improvement of +0.075 over the baseline, using fewer chunks than (c), indicating that the relevance of the additional chunks selected through the citation-graph structure matters more than quantity alone.

Although condition (b) shows the best performance among the three conditions, its recall of 0.948 still leaves a gap of 0.052 from the theoretical ceiling of 1.000. This gap does not necessarily indicate a system failure; rather, it more plausibly reflects articles that are textually and structurally difficult to reach. For example, articles with a thematic or

implicit relationship that is not expressed through an explicit article-number reference, as discussed further in the stratified analysis below.

C. Single-Hop vs. Multi-Hop Stratified Analysis

The results in Table III show that the recall and Full-Hit Rate differences between conditions are not evenly distributed. The entire improvement shown by condition (b) is concentrated in the multi-hop query group, while the single-hop group is unaffected. This pattern is consistent with the design of the graph-expansion mechanism, which is inherently active only when a retrieved article has an explicit-reference connection to another article.

TABLE III
SINGLE-HOP AND MULTI-HOP STRATIFICATION RESULTS

Condition	Single-hop recall/fh	Multi-hop recall/fh
(a)	1.000/1.000	0.808/0.600
(b)	1.000/1.000	0.896/0.750
(c)	1.000/1.000	0.821/0.600

For the single-hop group, all three conditions achieve perfect performance (recall = 1.000, Full-Hit Rate = 1.000). This confirms two things: first, that the hybrid RRF component alone is sufficient to handle queries whose answers are contained within a single article; and second, that graph expansion in condition (b) introduces no disruption for queries that do not require cross-referencing.

For the multi-hop group, condition (b) achieves a recall of 0.896 and a Full-Hit Rate of 0.750, an improvement of 0.088 in recall and 0.150 in Full-Hit Rate over baseline (a). Condition (c), with 20 chunks, achieves only a recall of 0.821 and a Full-Hit Rate of 0.600. Nearly identical to the baseline despite carrying twice as many chunks. This gap between (b) and (c) in the multi-hop group is empirical evidence that the improvement shown by condition (b) is not merely a context quantity effect, but an effect of the selectivity of chunk selection achieved through the graph structure.

D. Has-Edge vs. No-Edge Stratified Analysis

Further stratification divides the 20 multi-hop queries into two subgroups based on the presence of an explicit edge in the citation graph: 14 queries with a graph connection and 6 queries without one. Full results are presented in Table IV.

TABLE IV
HAS EDGE AND NO EDGE STRATIFICATION RESULTS

Stratum	n	Condition	Recall	Full-hit
Single-hop	20	(a)	1.000	1.000
		(b)	1.000	1.000
		(c)	1.000	1.000
Multi-hop+edge	14	(a)	0.762	0.500
		(b)	0.887	0.714
		(c)	0.780	0.500
Multi-hop-no-edge	6	(a)	0.917	0.833
		(b)	0.917	0.833
		(c)	0.917	0.833

Three main patterns emerge from the results in Table IV, collectively providing a strong causal argument. First, for multi-hop queries with edges, condition (b) improves recall by 0.125 and the Full-Hit rate by 0.214 compared to baseline (a). In contrast, condition (c), which expands retrieval to 20 chunks, yields a marginal recall improvement of only 0.018 and no change in the Full-Hit rate. The performance gap between conditions (b) and (c) within this stratum, despite condition (c) retaining a larger candidate pool, provides robust evidence that additional chunks retrieved via the citation graph possess higher relevance than those retrieved through pure Reciprocal Rank Fusion (RRF) with a lower threshold. Individual query cases clearly illustrate this behavior. For instance, a query regarding illegal access to an electronic system (targeting Articles 30 and 46) achieved only 50% recall under the baseline because it omitted Article 46, the sanction provision. Citation-graph expansion correctly retrieved Article 46 through its explicit reference from Article 30, raising the recall to 100%. The same mechanism similarly resolved queries related to phishing (targeting Articles 35 and 51) and criminal facilitation (targeting Articles 34 and 50), with recall improving from 50% to 100% in both cases. In each instance, the provision missed under the baseline was specifically the sanction article linked to an already retrieved prohibition article.

The second pattern, for multi-hop queries without edges, all three conditions produced identical results, achieving a recall of 0.917 and a Full-Hit rate of 0.833. Graph expansion offered neither benefit nor harm when candidate articles lacked explicit connecting edges. This indicates that the expansion mechanism operates selectively, activating only when supported by the underlying graph structure without indiscriminately introducing noise into the retrieval results. A representative case demonstrates this operational limit: for a query on jurisdictional disputes in international electronic transactions (targeting Articles 2 and 18), no citation edge connects the two provisions. Consequently, graph expansion failed to surface Article 2 regardless of the condition, leaving recall at 50% across both conditions (a) and (b). This limitation stems from the absence of an explicit cross-reference in the source text rather than insufficient search depth, aligning with the ablation finding that deeper two-hop traversals cannot compensate for a missing edge.

The third pattern, cross-condition symmetry in the single-hop and no-edge strata: identical performance in these two strata shows that the difference observed in the multi-hop+edge stratum is not an artefact of query-distribution differences or other uncontrolled characteristics. The variation is isolated precisely in the stratum theoretically predicted to benefit from the expansion mechanism.

Overall, this stratification pattern is consistent with the hypothesis that citation-graph expansion is selectively effective: the mechanism works if, and only if, an explicit inter-article reference structure is available in the corpus, and has no effect outside that condition.

E. Additional Ablation: Expansion Depth, Edge Direction, and Candidate Cap

TABLE V
ADDITIONAL ABLATION RESULT

Variant	Recall Overall	Full-Hit Overall	Recall Multi-hop	Full-Hit Multi-hop
Graph baseline (1-hop, cap=5, bidirectional)	0.948	0.875	0.896	0.750
2-hop	0.948	0.875	0.896	0.750
Cap = 3	0.935	0.850	0.871	0.700
Cap = 8	0.969	0.925	0.938	0.850
Forward-only edge	0.904	0.800	0.808	0.600
Reverse-only edge	0.969	0.925	0.938	0.850

To evaluate the optimal design choices of the graph mechanism, three additional ablation experiments were conducted: expansion depth (1-hop vs. 2-hop), edge direction (forward vs. reverse), and candidate cap (3, 5, and 8). Two-hop expansion yielded results identical to the 1-hop configuration, recording a recall of 0.948 and a Full-Hit rate of 0.875. This outcome indicates that citation chains extending beyond a single reference provide no incremental retrieval benefit within this corpus. In terms of edge direction, reverse expansion—retrieving a sanction article based on an already retrieved prohibition article—matched the top-performing candidate cap of 8, with a recall of 0.969 and a Full-Hit rate of 0.925. Conversely, forward expansion, which retrieves prohibition articles from sanction articles, reproduced the baseline without graph expansion (recall = 0.904, Full-Hit rate = 0.800). This asymmetry demonstrates that the performance gains of the citation graph derive almost entirely from the reverse direction. This finding aligns with the corpus characteristics discussed in Section Has-Edge vs. No-Edge Stratified Analysis, where queries more frequently retrieve prohibition articles directly through semantic or lexical similarity, leaving linked sanction articles as the primary retrieval gap. Finally, increasing the candidate cap from 5 to 8 improved the multi-hop Full-Hit rate to 0.850. This improvement suggests that the default cap of 5 may be overly conservative, though higher caps inherently increase context length during the generation stage.

F. Statistical Testing

Comparisons between conditions were tested using the paired Wilcoxon signed-rank test. Full results are presented in Table VI.

TABLE VI
STATISTICAL TEST RESULT

Stratum	Comparison	W	p	r	n_eff
Single-hop (n=20)	(a) vs (b)	-	-	-	-
Single-hop (n=20)	(b) vs (c)	-	-	-	-
Multi-hop all (n=20)	(a) vs (b)	0.0	0.059	1.000	4
Multi-hop all (n=20)	(b) vs (c)	1.5	0.098	0.900	5
Multi-hop+edge (n=14)	(a) vs (b)	0.0	0.059	1.000	4
Multi-hop+edge (n=14)	(b) vs (c)	1.5	0.098	0.900	5
Multi-hop-no-edge (n=6)	(a) vs (b)	-	-	-	-
Multi-hop-no-edge (n=6)	(b) vs (c)	-	-	-	-

To quantify the uncertainty around the point estimates reported in Table II and Table III, 95% confidence intervals were computed via bootstrap resampling (10,000 resamples), presented in Table VII.

TABLE VII
BOOTSTRAP 95% CI

Condition	Subset	Recall (95% CI)	Full-Hit (95% CI)
(a) Baseline	Overall	0.904 (0.842–0.963)	0.800 (0.675–0.925)
(a) Baseline	Multi-hop	0.808 (0.708–0.908)	0.600 (0.400–0.800)
(b) Graph	Overall	0.948 (0.898–0.988)	0.875 (0.775–0.975)
(b) Graph	Multi-hop	0.896 (0.808–0.975)	0.750 (0.550–0.900)
(c) Control	Overall	0.910 (0.850–0.963)	0.800 (0.675–0.925)
(c) Control	Multi-hop	0.821 (0.721–0.917)	0.600 (0.400–0.800)

The test could not be performed for the single-hop and multi-hop-no-edge strata because every pair produced an identical recall difference between conditions ($n_{\text{eff}} = 0$). In the single-hop stratum, all three conditions achieve perfect recall (1.000), leaving no variation to test. In the multi-hop-no-edge stratum, all three conditions likewise produce identical recall (0.917), confirming that graph expansion has no effect, positive or negative, on queries whose articles are not connected via an explicit edge in the citation graph.

Identical test results were obtained for multi-hop all and multi-hop+edge, since every pair with a non-zero difference originates from the `has_edge = True` subgroup. For the (a) vs. (b) comparison, $W = 0.0$, $p = 0.059$, with a rank-biserial effect

size of $r = 1.000$ at $n_{\text{eff}} = 4$. For the (b) vs. (c) comparison, $W = 1.5$, $p = 0.098$, with $r = 0.900$ at $n_{\text{eff}} = 5$. Neither p -value reaches significance at $\alpha = 0.05$.

This non-significance must be interpreted with caution, as it is primarily driven by the limited effective sample size n_{eff} . Given that the theoretical minimum required for a two-sided Wilcoxon signed-rank test to achieve statistical significance $p < 0.05$ is $n_{\text{eff}} \geq 6$, the available statistical power is inherently insufficient to detect even substantively large effects. Notably, the rank-biserial correlation of $r = 1.000$ for the comparison between conditions (a) and (b) indicates that every non-tied pair consistently favored condition (b) without exception, representing the maximum attainable effect size under the Wilcoxon test. Similarly, the rank-biserial correlation of $r = 0.900$ for the comparison between conditions (b) and (c) constitutes a large effect size according to Cohen's conventions.

This limitation in statistical power is a direct consequence of the limited size of the evaluation dataset. Of the 14 multi-hop+edge queries, only a small subset produced a recall difference between conditions. Most queries produced identical recall across all conditions and therefore did not contribute to the test. Replication with a larger dataset, particularly with more multi-hop queries spanning a wider range of ground-truth difficulty, is needed to obtain adequate statistical confirmation.

Nonetheless, the convergence of three strands of empirical evidence strengthens the plausibility of the graph-expansion effect, even though the results remain suggestive rather than statistically significant, and larger-scale replication is needed before a claim of effectiveness can be firmly established: (1) a large, consistently directional effect size in the stratum where it was predicted; (2) no effect in the strata predicted to be unaffected; and (3) the quantity-control condition (c) failing to replicate condition (b)'s improvement despite carrying more chunks, isolating the effect of graph selectivity from that of quantity alone.

G. Discussion

The main finding of this study is that citation-graph expansion consistently improves retrieval performance for queries requiring cross-article synthesis, with a Full-Hit Rate improvement of +0.150 in the multi-hop+edge group relative to baseline. This improvement cannot be explained by chunk quantity alone: the quantity-control condition (c), which carries 34% more chunks through pure RRF, yields a Full-Hit Rate improvement of only +0.000 in the same stratum. This gap indicates that the relevance of chunks selected through the graph structure matters more than broader retrieval coverage.

This result is theoretically consistent with the structural characteristics of Indonesian legislation. UU ITE, like most criminal law, separates prohibition articles and sanction articles into distinct sections that are linearly distant from one another yet connected through explicit references. Retrieval systems based on semantic and lexical similarity are not designed to traverse this kind of structural relationship: a

query about the sanction for a particular violation is, semantically, more similar to the relevant prohibition article than to the sanction article that actually contains the sought information. Reverse expansion in the citation graph addresses this gap directly by pulling in the sanction article based on an explicit reference relationship rather than textual similarity.

The performance improvement is concentrated exclusively in multi-hop+edge queries, while single-hop and multi-hop-no-edge queries are unaffected. This confirms that the system's mechanism operates selectively, as designed, rather than as a random side effect active only when the graph structure supports meaningful expansion, and not burdening simple queries with unnecessary added complexity. However, this selectivity also reveals a limitation of the system. The six multi-hop-no-edge queries show identical performance across all three conditions, indicating that some inter-article relationships are implicit or thematic rather than expressed through an explicit article reference, and therefore require a different approach such as a knowledge graph built on legal entities or a re-ranking model capable of understanding normative hierarchy.

Beyond retrieval mechanics, the reliability of the underlying legal text itself warrants careful consideration. Because Law No. 11 of 2008 on Information and Electronic Transactions (UU ITE) has undergone two amendments, in 2016 and 2024, the consolidated corpus represents exclusively the current in-force text of each provision, with amendment status logged per article for traceability, as detailed in Section Corpus and Preprocessing. However, the current system lacks an automated mechanism to flag provisions that may be superseded by amendments beyond 2024. Furthermore, it does not verify article text against officially gazetted sources (*Lembaran Negara*), which presents an important limitation given that the corpus was digitized via optical character recognition (OCR) without systematic official cross-verification. Consequently, the proposed architecture should be positioned as a research prototype for legal information retrieval and question answering rather than a substitute for professional legal advice, pending broader validation against official legal repositories and rigorous quality assessment of generated responses.

IV. CONCLUSION

This study proposes and evaluates a Retrieval-Augmented Generation system for UU ITE legal documents that combines Reciprocal Rank Fusion-based hybrid retrieval with inter-article citation-graph expansion. The study is motivated by a structural characteristic of Indonesian legislation, in which prohibition and sanction articles are placed in separate sections yet connected through explicit references, creating a multi-hop retrieval challenge that purely semantic-lexical similarity cannot resolve. The primary contribution of this study is a staged ablation design that isolates the specific effect of citation-graph selectivity from retrieval-depth and

candidate-quantity factors, thereby addressing a critical validation gap in prior graph-based RAG and Indonesian legal retrieval research.

Evaluation on a 63-article corpus (spanning three regulatory layers) with 40 query ground-truth pairs shows that the condition with citation-graph expansion is associated with a recall of 0.948 and a Full-Hit Rate of 0.875, higher than both the baseline (recall 0.904; Full-Hit Rate 0.800) and the quantity-control condition (recall 0.910; Full-Hit Rate 0.800), despite the latter carrying 34% more chunks. All improvement is selectively concentrated in multi-hop queries with an explicit graph connection (Full-Hit Rate rising from 0.500 to 0.714), with no effect on single-hop queries or on multi-hop queries lacking a graph connection.

The Wilcoxon test on this stratum yields a large effect size ($r = 1.000$, $p = 0.059$) that does not reach significance at $\alpha = 0.05$, owing to limited statistical power from the small sample. Given that ground-truth validation was performed by a single researcher, and that a formal evaluation of generated-answer quality has not yet been conducted, this study identifies several directions for future work: developing an integrated scoring scheme, extending reference extraction, and testing generalisability on other Indonesian legislation with a similar inter-article reference structure.

REFERENCES

- [1] N. Pipitone and G. H. Alami, "LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain," Aug. 2024.
- [2] R. Kalra *et al.*, "HyPA-RAG: A Hybrid Parameter Adaptive Retrieval-Augmented Generation System for AI Legal and Policy Applications," *1st Work. Cust. NLP Prog. Challenges Cust. NLP a Domain, Appl. Group, or Individ. Cust. 2024 - Proc. Work.*, pp. 237–256, Feb. 2024, doi: 10.18653/v1/2024.customnlp4u-1.18.
- [3] M. Reuter *et al.*, "Towards Reliable Retrieval in RAG Systems for Large Legal Datasets," in *Proceedings of the Natural Language Processing Workshop 2025*, Suzhou, China: Association for Computational Linguistics, 2025, pp. 17–30. doi: 10.18653/v1/2025.nllp-1.3.
- [4] H. De Martim, "An Ontology-Driven Graph RAG for Legal Norms: A Hierarchical, Temporal, and Deterministic Approach," *Front. Artif. Intell. Appl.*, vol. 416, pp. 282–287, 2025, doi: 10.3233/FAIA251598.
- [5] K. Choudhary, T. Kandoi, and S. Patel, "Hybrid GraphRAG for Cross-Lingual Legal Citation Retrieval: A Multi-Signal Fusion Approach for Swiss Legal Information Systems," vol. 15, no. 04, pp. 28374–28379, 2026, doi: 10.18535/ijecs/v15i03.5461.
- [6] A. Y. Yulestiono, A. P. Sari, and A. Rakhmadi, "Evaluasi Hybrid Retrieval Berbasis Reciprocal Rank Fusion untuk Pencarian Informasi pada Dokumen Peraturan Daerah," *Pros. Semin. Nas. Inform. Bela Negara*, vol. 6, no. 1, pp. 148–153, Jul. 2026, doi: 10.33005/santika.v6i1.1233.
- [7] A. Z. Pratama, A. Marinta, B. Triyudanto, M. Saman, and T. Noor, "Retrieval-Augmented Generation for Indonesian Criminal Law Information Using the LLaMA Model," *Innov. Informatics Artif. Intell. Res.*, vol. 1, no. 1, pp. 35–41, 2025, [Online]. Available: <https://journal.itk.ac.id/index.php/IIAIR/article/view/1306>
- [8] A. F. Ferraris, D. Audrito, G. Siragusa, and A. Piovano, "Legal Chunking: Evaluating Methods for Effective Legal Text Retrieval," *Front. Artif. Intell. Appl.*, vol. 395, pp. 275–281, 2024, doi: 10.3233/FAIA241255.
- [9] Z. Zhong, H. Liu, X. Cui, X. Zhang, and Z. Qin, "Mix-of-

- Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation,” *Proc. - Int. Conf. Comput. Linguist. COLING*, pp. 5756–5774, Jan. 2025.
- [10] Muhammad Rafli Aditya H., Muhammad Ilham, Dewi Fatmarani Surianto, and Abdul Muis Mappalotteng, “Evaluasi Pengukuran Semantik Sinonim KBBI Menggunakan Pendekatan Word Embedding,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 14, no. 2, pp. 112–120, May 2025, doi: 10.22146/jnteti.v14i2.17117.
- [11] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “NormFace: L2 hypersphere embedding for face verification,” in *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, New York, NY, USA: ACM, Oct. 2017, pp. 1041–1049. doi: 10.1145/3123266.3123359.
- [12] H. Steck, C. Ekanadham, and N. Kallus, “Is Cosine-Similarity of Embeddings Really About Similarity?,” in *Companion Proceedings of the ACM Web Conference 2024 (WWW ’24)*, New York, NY, USA: Association for Computing Machinery, 2024, pp. 887–890. doi: 10.1145/3589335.3651526.
- [13] J. Johnson, M. Douze, and H. Jegou, “Billion-Scale Similarity Search with GPUs,” *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.
- [14] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*, vol. 3, no. 4. 2009. doi: 10.1561/1500000019.
- [15] Y. Setiawan and L. A. Wulandhari, “Comparative Analysis of IndoBERT and LSTM for Multi-Label Text Classification of Indonesian Motivation Letter,” *J. Online Inform.*, vol. 10, no. 2, pp. 260–269, 2025, doi: 10.15575/join.v10i2.1499.
- [16] M. Siino, I. Tinnirello, and M. La Cascia, “Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers,” *Inf. Syst.*, vol. 121, no. March 2023, p. 102342, 2024, doi: 10.1016/j.is.2023.102342.
- [17] J. Pardede and D. Darmawan, “Perbandingan Algoritma Stemming Porter, Sastrawi, Idris, Dan Arifin & Setiono Pada Dokumen Teks Bahasa Indonesia,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 12, no. 1, pp. 69–76, 2025, [Online]. Available: <https://jtiik.ub.ac.id/index.php/jtiik/article/view/8860>
- [18] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, “Reciprocal rank fusion outperforms condorcet and individual rank learning methods,” *Proc. - 32nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 2009*, pp. 758–759, 2009, doi: 10.1145/1571941.1572114.
- [19] X. Zhu, Y. Xie, Y. Liu, Y. Li, and W. Hu, “Knowledge Graph-Guided Retrieval Augmented Generation,” *Proc. 2025 Annu. Conf. Nations Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Long Pap. NAACL-HLT 2025*, vol. 1, pp. 8912–8924, 2025, doi: 10.18653/v1/2025.naacl-long.449.
- [20] M. Hei, Q. Liu, S. Zhang, H. Shi, and J. Duan, “A Heterogeneous Graph Based on Legal Documents and Legal Statute Hierarchy for Chinese Legal Case Retrieval,” *IEEE Access*, vol. PP, p. 1, 2024, doi: 10.1109/ACCESS.2024.3419139.
- [21] Z. Shen *et al.*, “GEAR: Graph-enhanced Agent for Retrieval-augmented Generation,” *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 12049–12072, 2025, doi: 10.18653/v1/2025.findings-acl.624.
- [22] N. S. Muninggar and A. A. Krisnadhi, “LexID: The Metadata and Semantic Knowledge Graph Construction of Indonesian Legal Document,” *J. Ilmu Komput. dan Inf.*, vol. 16, no. 1, pp. 15–46, Mar. 2023, doi: 10.21609/jiki.v16i1.1096.
- [23] A. Ghassel *et al.*, *Hierarchical Lexical Graph for Enhanced Multi-Hop Retrieval*, vol. 2, no. 1. Association for Computing Machinery, 2025. doi: 10.1145/3711896.3737233.
- [24] S. Huang *et al.*, “DriveLegal : Toward legally compliant driving via trustworthy hybrid retrieval-augmented LLMs,” *Expert Syst. Appl.*, vol. 314, no. November 2025, p. 131593, 2026, doi: 10.1016/j.eswa.2026.131593.
- [25] N. F. Liu *et al.*, “Lost in the Middle: How Language Models Use Long Contexts,” *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 157–173, Feb. 2024, doi: 10.1162/tacl_a_00638.
- [26] B. Peng *et al.*, “Graph Retrieval-Augmented Generation: A Survey,” *ACM Trans. Inf. Syst.*, vol. 44, no. 2, 2026, doi: 10.1145/3777378.
- [27] D. Abts *et al.*, “Think Fast: A Tensor Streaming Processor (TSP) for Accelerating Deep Learning Workloads,” in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, IEEE, May 2020, pp. 145–158. doi: 10.1109/ISCA45697.2020.00023.
- [28] A. Grattafiori *et al.*, “The Llama 3 Herd of Models,” Nov. 2024, doi: 10.48550/arXiv.2407.21783.
- [29] S. Xia *et al.*, “Ground Every Sentence: Improving Retrieval-Augmented LLMs with Interleaved Reference-Claim Generation,” in *2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Proceedings of the Conference Findings, NAACL 2025*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 969–988. doi: 10.18653/v1/2025.findings-naacl.55.
- [30] J. Lee, D. Kim, S. Hwang, H. Kim, and G. G. Lee, “KOBLEX: Open Legal Question Answering with Multi-hop Reasoning,” *EMNLP 2025 - 2025 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 4019–4053, 2025, doi: 10.18653/v1/2025.emnlp-main.200.
- [31] Z. Yang *et al.*, “Hotpotqa: A dataset for diverse, explainable multi-hop question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 2369–2380. doi: 10.18653/v1/d18-1259.
- [32] H. Wang *et al.*, “A Query-Driven Graph Retrieval Framework with Adaptive Pruning for Multi-Hop Question Answering,” *Electron.*, vol. 15, no. 6, pp. 1–22, 2026, doi: 10.3390/electronics15061263.
- [33] A. Salemi, “Evaluating Retrieval Quality in Retrieval-Augmented Generation,” pp. 2395–2400, 2024, doi: 10.1145/3626772.3657957.
- [34] Y. Tang and Y. Yang, “MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries,” in *Conference on Language Modeling (COLM 2024)*, 2024. doi: 10.48550/arXiv.2401.15391.
- [35] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bull.*, vol. 1, no. 6, p. 80, Dec. 1945, doi: 10.2307/3001968.
- [36] J. Demšar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
- [37] D. S. Kerby, “The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation,” *Compr. Psychol.*, vol. 3, p. 11.IT.3.1, 2014, doi: 10.2466/11.it.3.1.