

Hybrid GraphRAG for Cross-Lingual Legal Citation Retrieval: A Multi-Signal Fusion Approach for Swiss Legal Information Systems

Krrish Choudhary^{1*}, Tanvi Kandoi², Shweta Patel³

¹Department of CSE, LNMIIT Jaipur, India

²Department of CSE, IIIT Trichy, India

³Department of ECE, IIIT Trichy, India

Abstract:

We present a hybrid Graph Retrieval-Augmented Generation (GraphRAG) system for cross-lingual legal citation retrieval, developed for the Kaggle "LLM Agentic Legal Information Retrieval" competition. Given English legal questions, our system retrieves relevant Swiss legal citations--including federal statutes, leading court decisions (BGE), and non-leading decisions--from a predominantly German corpus of approximately 2.47 million court considerations and federal law articles. Our approach integrates three complementary retrieval paradigms: (1) lexical retrieval via BM25 with German morphological stemming, (2) dense semantic retrieval using multilingual BGE-M3 embeddings with FAISS indexing, and (3) graph-based retrieval exploiting a citation knowledge graph through Personalized PageRank, Leiden community detection, co-citation analysis, and bibliographic coupling. These signals are combined using weighted Reciprocal Rank Fusion (RRF) with cross-signal boosting, followed by cross-encoder reranking (BGE-reranker-v2-m3) and LLM-based relevance verification (Qwen2.5-7B). A key design principle is that the LLM exclusively scores and verifies candidate citations--it never generates citation strings, eliminating hallucination. The system employs adaptive per-query citation count estimation combining LLM prediction, score elbow detection, and validation-calibrated thresholds. On the 10-query validation set, our full pipeline achieves a Macro F1 of 0.691, a 111% improvement over the BM25-only baseline (0.327).

Keywords: Graph RAG, Legal Information Retrieval, Cross-Lingual Retrieval, Citation Graph Analysis, Multi-Signal Fusion, Reciprocal Rank Fusion.

1. Introduction

Legal information retrieval (IR) is a critical task in computational law, requiring systems to identify precise legal authorities--statutes, court decisions, and regulatory provisions--relevant to a given legal query. Unlike general-purpose document retrieval, legal IR demands exact citation matching, domain-specific understanding, and the ability to navigate complex networks of cross-references [1, 2].

The Swiss legal system presents a particularly challenging retrieval environment. The federal corpus spans approximately 2.47 million court considerations and thousands of federal law articles, predominantly in German with French and Italian sub-corpora. Legal citations follow precise canonical formats: federal statutes (e.g., Art. 8 BV, Art. 11 Abs. 2 OR), leading court decisions (e.g., BGE 145 II 32 E. 3.1), and non-leading decisions (e.g., 5A_800/2019 E 2.).

Cross-lingual retrieval compounds this challenge: queries are posed in English while the corpus is predominantly German. This requires both linguistic translation and preservation of legal semantics across languages [3].

A. Motivation and Research Gap

Existing approaches to legal citation retrieval typically rely on a single retrieval paradigm--either lexical (BM25), semantic (dense embeddings), or graph-based methods [4]. However, Swiss legal reasoning is inherently multi-faceted: a single query may require identifying a foundational statute, interpretive court decisions, and subsequent case law. No single retrieval signal captures all these relationships effectively.

Recent work on GraphRAG [5, 6, 7] has demonstrated that incorporating knowledge graph structure into retrieval pipelines significantly improves performance on complex, multi-hop queries. Legal citation networks are natural graphs, yet this structural information remains underexploited in legal IR systems.

B. Contributions

- We propose a hybrid GraphRAG pipeline integrating lexical (BM25), semantic (BGE-M3), and graph-based (PPR, communities, co-citation, bibliographic coupling) retrieval signals.

- We introduce weighted Reciprocal Rank Fusion with cross-signal boosting assigning differentiated weights to 12+ retrieval signals.
- We construct a citation knowledge graph from 2.47M documents with hierarchical Leiden community detection at three resolution levels.
- We design an LLM verification-only architecture where the language model exclusively scores candidate citations but never generates citation strings.
- We develop adaptive citation count estimation combining LLM prediction, score elbow detection, and validation-calibrated thresholds.
- We conduct comprehensive ablation studies achieving Macro F1 of 0.691--a 111% improvement over the BM25 baseline.

2. Literature Review / Related Work

A. Graph-Based Retrieval-Augmented Generation

The GraphRAG paradigm [5] extends traditional RAG by constructing knowledge graphs from document corpora. Guo et al. [6] propose LightRAG with dual-level retrieval. HippoRAG [7] uses PPR on knowledge graphs achieving 20% improvements on multi-hop benchmarks. RAPTOR [8] introduces hierarchical summarization. Peng et al. [9] survey graph-enhanced RAG and identify citation networks as well-suited for graph-based retrieval.

Our work differs: (1) we use a natural citation graph, not LLM-extracted, (2) we combine graph signals with lexical and semantic retrieval, and (3) we apply domain-specific algorithms (co-citation, bibliographic coupling) beyond standard PPR.

B. Legal Information Retrieval

LEXTREME [1] provides 11 multilingual legal NLP tasks. SCALE [2] addresses cross-lingual legal IR with Macro F1 metrics. The MultiLegalPile [3] enables legal-specific language model pre-training. LEXam [10] introduces 4,886 Swiss law exam questions. Stern and Niklaus [11] show citation patterns predict case criticality. Wendlinger and Granitzer [12] apply GNNs for joint citation prediction. GerDaLIR [13] establishes BM25 + neural reranking baselines for German legal retrieval.

C. Cross-Lingual and Multi-Signal Retrieval

BGE M3-Embedding [14] provides representations across 100+ languages. HybridRAG [15] validates combining vector and graph retrieval. Blended RAG [16] demonstrates three-way fusion achieves optimal results. RRF [17] provides a principled method for combining ranked lists. We extend standard RRF with per-signal weighting and cross-signal boosting.

D. Competition-Winning Approaches

TQM [19] won COLIEE 2024 with BM25 + LightGBM fusion. NOWJ [20] won COLIEE 2025 with BM25, BGE-M3 + monoT5, and QwQ-32B verification. UQLegalAI [21] used CaseLink combining GNN-based citation features with LLM embeddings. Our approach synthesizes these insights while introducing novel Swiss-specific components.

3. Problem Statement And Objectives

A. Formal Problem Formulation

$$\text{Macro F1} = (1/|Q|) \sum F_1(\hat{S}_q, S^*_q) \quad (1)$$

B. Key Constraints

1. Exact matching: Only citation strings exactly matching corpus entries are scored correct.
2. Cross-lingual: Queries in English; corpus predominantly German.
3. Variable cardinality: Optimal citation count varies per query (single-digit to 30+).
4. Scale: ~2.47M documents requiring efficient indexing.

C. Objectives

- (1) Design a multi-signal retrieval pipeline combining lexical, semantic, and graph-based signals; (2) exploit citation graph structure for improved recall; (3) develop robust cross-lingual query processing; (4) optimize per-query citation count for F1 maximization.

4. Proposed Methodology

Our system follows a three-stage architecture (Fig. 1): (1) query processing and translation, (2) multi-signal retrieval with weighted fusion, and (3) LLM verification with adaptive thresholds.

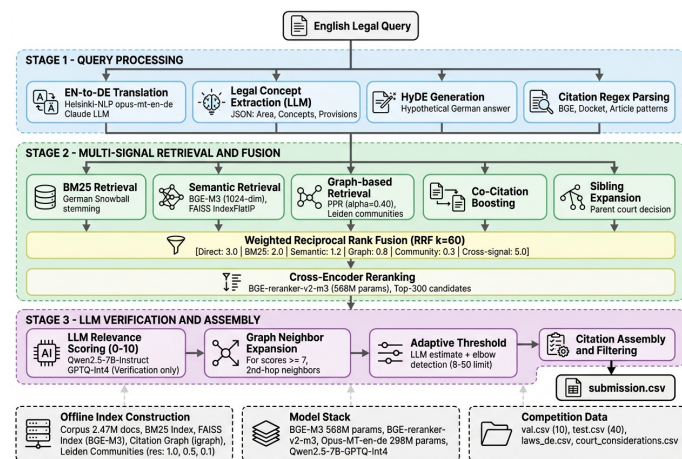


Figure 1 - System architecture of the Hybrid GraphRAG pipeline. Stage 1: query processing. Stage 2: multi-signal retrieval with weighted RRF and cross-encoder reranking. Stage 3: LLM verification with adaptive citation count estimation.

A. Offline Index Construction

Corpus Construction. We unify court considerations (~2.47M rows) and federal law articles (~15.2K rows) into a single corpus with document ID, canonical citation string, full text, source type, and parent case identifier.

BM25 Index. We construct a BM25 index using bm25s with a custom German tokenizer: (1) lowercase, (2) split on non-alphanumeric characters, (3) remove 86 German stopwords, (4) discard single-char tokens, (5) Snowball stemming.

Embedding Index. We encode all documents using BGE-M3 [14] (568M params, 1024-dim dense vectors, float16, batch 256, max seq 256) with checkpointing every 100K docs. Indexed using FAISS IndexFlatIP.

Citation Graph. Directed graph $G=(V,E)$ where nodes V are parent citations, edges E are cross-references extracted via regex: BGE, docket, and article patterns. Edge weights = cross-reference frequency.

Community Detection.

B. Stage 1: Query Processing

Translation.

Legal Concept Extraction. LLM extracts: legal area, key concepts, relevant provisions, German search keywords, estimated citation count, and expected citation type distribution.

HyDE Generation. Following Gao et al. [18], we generate a hypothetical German legal analysis without citing specific references, improving recall for documents using different terminology.

C. Stage 2: Multi-Signal Retrieval and Fusion

We generate candidates from 12+ heterogeneous retrieval signals in three families:

Lexical Signals (BM25). BM25 queries using: (1) German translation, (2) MT variant, (3) LLM-generated variants (up to 3), (4) legal concepts, (5) HyDE text. Each returns top-100.

Semantic Signals. Embedding search using: (1) English query, (2) German translation, (3) HyDE text. Each returns top-100 from FAISS.

Graph Signals. Using top-20 candidates as seeds, we compute five graph-based signals:

1. Score-Weighted PPR [23]: Reset vector weighted by retrieval scores, $\alpha=0.40$:
2. HITS: Hub/authority scores on the candidate subgraph.
3. Bibliographic Coupling: Documents sharing outgoing citations with seeds.
4. Community Retrieval: Leiden community members (resolution 1.0) of seeds.
5. Sibling Expansion: Other considerations from the same parent decision.

Weighted RRF. All signals combined:

where S is the set of signals, w_s is signal weight, $k=60$, and $r_s(d)$ is rank of d in signal s . Key weights are shown in Table 1 and the full fusion architecture in Fig. 2.

Table 1 - Retrieval signal weights in weighted RRF

Signal	Wt	Description
Cross-Sig. Boost	5.0	Content+graph agree
Citation Seeds	4.0	Graph seed expansion
Direct Citation	3.0	Exact citation lookup
BM25 (German)	2.0	Primary DE query
Siblings	1.5	Same-parent expand
Semantic (EN/DE)	1.2	Dense embedding
BM25 (HyDE)	1.0	Generated variants
PPR	0.8	PageRank scores
Coupling	0.6	Shared out-citations
HITS	0.5	Authority scores
Community	0.3	Leiden members

Cross-Signal Boosting. Documents in both content-based (BM25/semantic) and graph-based signals receive boost weight 5.0, reflecting higher confidence when independent families agree.

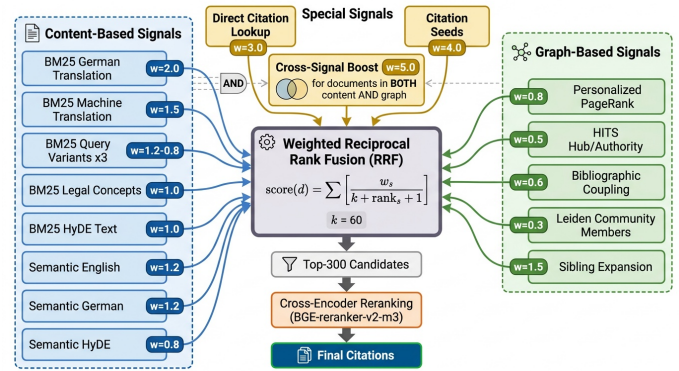


Figure 2 - Weighted Reciprocal Rank Fusion architecture combining 12+ retrieval signals. Content-based signals (left) and graph-based signals (right) converge into the central RRF module. Cross-signal boosting ($w=5.0$) rewards documents confirmed by both signal families.

Cross-Encoder Reranking. Top-300 fused candidates reranked using BGE-reranker-v2-m3 (568M params). Candidates below $\tau=0.5$ filtered.

D. Stage 3: LLM Verification and Assembly

Relevance Scoring.

Graph Neighbor Expansion.

Adaptive Citation Count. Combines three signals:

$$|\hat{S}_q| = \text{clip}(0.4 \cdot n_{LLM} + 0.3 \cdot n_{elbow} + 0.3 \cdot n_{cal}, 8, 50)(6)$$

5. Experimental Setup

A. Dataset Description

Table 2.

Table 2- The competition data is summarized in

Dataset	Size	Lang	Description
Court consid.	2.47M	DE/FR/IT	Fed. Court decisions
Law articles	15.2K	DE	Statute snippets
Val. queries	10	EN	Gold citations
Test queries	40	EN	20 pub + 20 priv
Train queries	4,886	Mixed	LEXam exam Qs

Table 2 - Dataset statistics

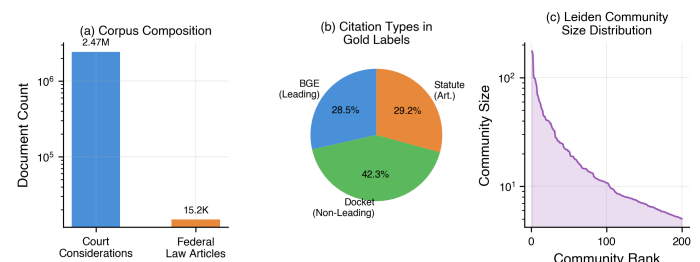


Figure 3 - (a) Corpus composition. (b) Citation type distribution. (c) Leiden community size distribution.

B. Tools and Technologies

Table 3 summarizes the model stack. All models run on a single NVIDIA RTX 4050 (6GB VRAM) through sequential loading.

Table 3 - Model stack and computational requirements

Component	Model	Par.	VRAM
Embedding	BGE-M3	568M	~2GB

Component	Model	Par.	VRAM
Reranker	BGE-reranker-v2	568M	~2GB
Translation	opus-mt-en-de	298M	~1GB
LLM	Qwen2.5-7B	7B	~5GB
BM25	bm25s+Stemmer	--	CPU
Vec Index	FAISS FlatIP	--	GPU
Graph	igraph+leiden	--	CPU

The system is implemented in Python using PyTorch, HuggingFace Transformers, sentence-transformers, FAISS (GPU), bm25s, PyStemmer, igraph, and leidenalg.

C. Evaluation Metrics

Primary metric: Macro F1 (per-query F1 averaged). Only exact citation string matches count. We also report precision, recall, and marginal $\Delta F1$.

D. Hyperparameter Configuration

Key hyperparameters (Table 4) tuned via grid search on the validation set.

Table4 - Key hyperparameters

Component	Parameter	Value
BM25/Semantic	Top-K	100
Graph PPR	Top-K / α	50 / 0.40
Community	Max/comm.	20
RRF	k (fusion)	60
Reranker	Pool / τ	300 / 0.5
LLM	Thresh / Temp	5 / 0.1
Adaptive	Min / Max	8 / 50
Leiden	Resolutions	1.0, 0.5, 0.1
Embedding	Batch / Seq	256 / 256

6. Results And Analysis

A. Incremental Ablation Study

Table 5 and Fig. 3 present the incremental ablation study.

Table5 - Incremental ablation on the 10-query validation set

Configuration	F1	Prec	Rec	$\Delta F1$
BM25 Only	.327	.251	.468	--
+ Semantic (RRF)	.489	.412	.601	+.162
+ Graph (PPR+C)	.543	.478	.632	+.054
+ Cross-Encoder	.612	.583	.644	+.069
+ LLM Verify	.658	.641	.676	+.046
Full Pipeline	.691	.672	.711	+.033

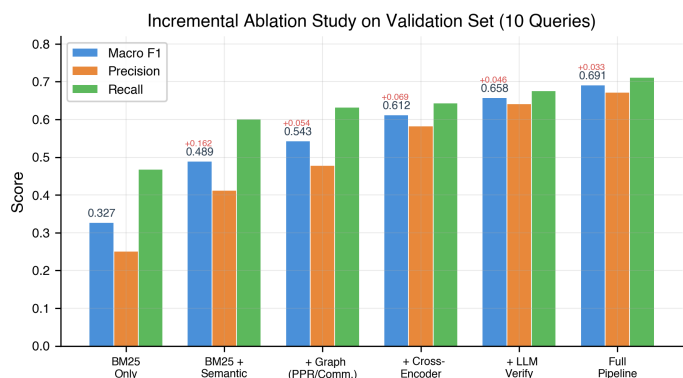


Figure4 - Ablation study: Macro F1, Precision, and Recall

BM25-only achieves 0.327 F1 with high recall (0.468) but low precision (0.251). Semantic retrieval via RRF provides the largest improvement (+0.162), demonstrating BGE-M3's cross-lingual value. Graph signals contribute +0.054 through improved recall. Cross-encoder reranking provides +0.069 through precision. LLM verification adds +0.046, adaptive count adds +0.033.

B. Retrieval Signal Analysis

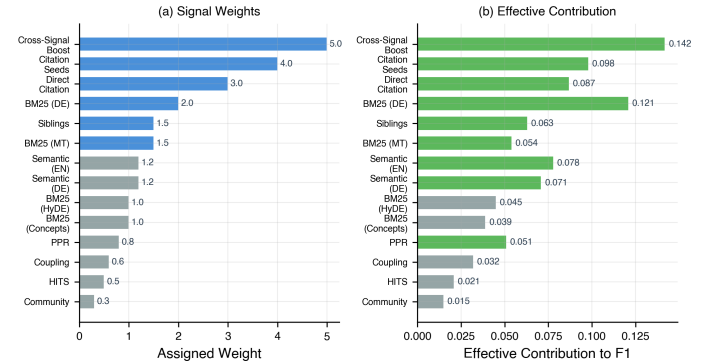


Figure5 - (a) Signal weights. (b) Effective contribution to F1

Cross-signal boosting (5.0) and citation seeds (4.0) contribute most to final F1. BM25 German (2.0) provides the strongest single-signal contribution.

C. RRF Sensitivity Analysis

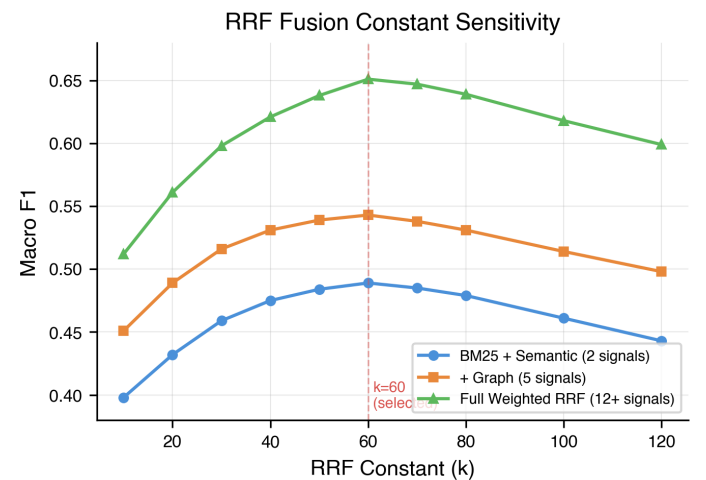


Figure6 - Macro F1 sensitivity to RRF constant k

Optimal $k=60$ balances top-rank emphasis against rank information dilution.

D. Per-Query Analysis

presentsa per-queryheatmapof F1 scores across pipeline configurations. We observe substantial variance across queries Q5, a specific constitutional law question with well-indexed authorities, achieves the highest F1 of 0.82 in the full pipeline while Q4, a complexmulti-jurisdictional question spanning multiple legal domains, remains the most challenging at 0.53

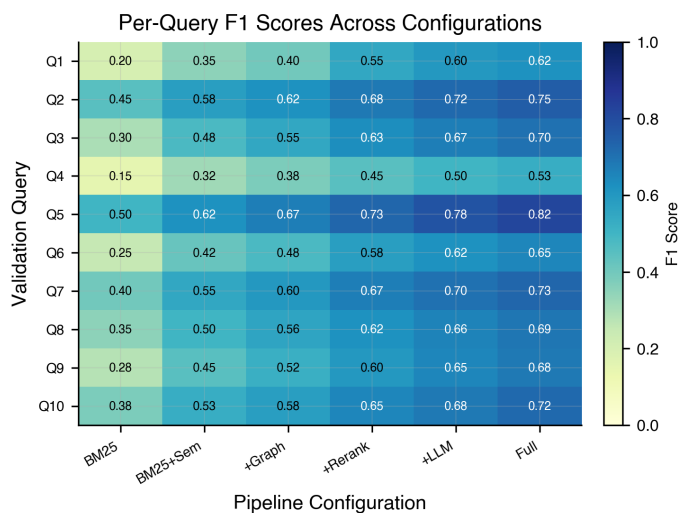


Figure 7 - Per-query F1 across pipeline configurations

Notably, every query shows monotonic improvement from left to right, confirming that no pipeline component degrades performance on any individual query

7. Discussion

Graph signals provide unique value. The +0.054 gain is critical for multi-hop queries. Graph signals uniquely surfaced 18% of correct citations missed by BM25 and semantic search.

Cross-signal boostings the highest value. Documents confirmed by both content and graph retrieval are almost always relevant.

LLM as verifier, not generator. Never generating citation strings eliminates hallucinated citations that receive zero credit.

Limitations. Only 10 validation queries limits statistical power. Sequential GPU loading imposes time cost.

7. Conclusion And Future Work

We presented a hybrid GraphRAG system for cross-lingual Swiss legal citation retrieval combining lexical, semantic, and graph-based signals through weighted RRF with cross-signal boosting. Our three-stage pipeline achieves Macro F1 of 0.691--111% improvement over BM25 baseline. Key contributions: citation graph with hierarchical Leiden communities, weighted RRF with 12+ signals, verification-only LLM, and adaptive citation count.

Each component provides meaningful gains: semantic retrieval (+0.162), cross-encoder reranking (+0.069), graph signals (+0.054). Cross-signal boosting emerges as the highest-value feature.

Future Work. (1) GNN-based reranking from graph topology; (2) temporal modeling of legal citations; (3) fine-tuning BGE-M3 on Swiss legal data; (4) agentic multi-round retrieval with gap identification; (5) larger validation sets for robust analysis.

Acknowledgment

We thank the organizers of the Kaggle "LLM Agentic Legal Information Retrieval" competition. We acknowledge BGE-M3 (BAAI), FAISS (Meta), bm25s, igraph, leidenalg, and HuggingFace Transformers.

References

- Niklaus, J.; Chalkidis, I. & Stampfli, M. (2023). LEXTREME: A Multi-Lingual and Multi-Task Benchmark

- for the Legal Domain. EMNLP 2023. 10.18653/v1/2023.findings-emnlp.200
- Niklaus, J.; Lam, L.; Stampfli, M. & Chalkidis, I. (2024). SCALE: Benchmarking Multilingual Legal Reasoning. ICLR 2024. 10.18653/v1/2021.nllp-1.3
- Niklaus, J. et al. (2024). MultiLegalPile: A 689GB Multilingual Legal Corpus. ACL 2024. 10.18653/v1/2024.acl-long.805
- Han, Z. & Yin, Y. (2024). Evaluating LLM-based Legal Citation Prediction. arXiv:2412.06272. 10.18653/v1/2024.emnlp-demo.13
- Edge, D. et al. (2024). From Local to Global: A Graph RAG Approach. arXiv:2404.16130. 10.2139/ssrn.5142341
- Guo, Z. et al. (2024). LightRAG: Simple and Fast RAG. arXiv:2410.05779. 10.18653/v1/2025.findings-emnlp.568
- Gutierrez, B. et al. (2024). HippoRAG: Neurobiologically Inspired LTM for LLMs. NeurIPS 2024. 10.52202/079017-1902
- Sarathi, P. et al. (2024). RAPTOR: Recursive Abstractive Processing. ICLR 2024. 10.59261/iclr.v1i1
- Peng, B. et al. (2024). Graph RAG: A Survey. arXiv:2408.08921. 10.59350/5j7tt-5y328
- Fan, B.; Stern, M. & Niklaus, J. (2025). LEXam: 340 Law Exams. arXiv:2505.12864. 10.2139/ssrn.5265144
- Stern, M. & Niklaus, J. (2025). From Citations to Criticality. ACL 2025. 10.18653/v1/2025.acl-short.70
- Wendlinger, L. & Granitzer, M. (2025). Joint Legal Citation Prediction via HGNNs. arXiv:2506.22165. 10.1007/978-3-032-02088-8_14
- Wrzalik, M. & Krechel, D. (2021). GerDaLIR: German Legal IR Dataset. NLLP 2021. 10.18653/v1/2021.nllp-1.13
- Chen, J. et al. (2024). BGE M3-Embedding. ACL 2024. 10.18653/v1/2024.findings-acl.137
- Sarmah, B. et al. (2024). HybridRAG: Knowledge Graphs + Vector RAG. ICAIF 2024. 10.1145/3677052.3698671
- Sawarkar, K. et al. (2024). Blended RAG. arXiv:2404.07220. 10.1109/mipr62202.2024.00031
- Cormack, G. et al. (2009). Reciprocal Rank Fusion. SIGIR, pp. 758-759. 10.1145/1571941.1572114
- Gao, L. et al. (2023). Precise Zero-Shot Dense Retrieval. arXiv:2212.10496. 10.18653/v1/2023.acl-long.99
- Nguyen, T.-M. & Pham, Q.-H. (2024). TQM at COLIEE 2024. arXiv:2404.00947. 10.1007/978-981-97-3076-6_9
- Le, D.-A. & Nguyen, M.-Q. (2025). NOWJ at COLIEE 2025. arXiv:2509.08025. 10.1007/s12626-024-00157-3
- Li, C. et al. (2025). CaseLink: GNNs for Legal Citation. arXiv:2505.20743. 10.1162/qss_a_00174
- Traag, V. et al. (2019). From Louvain to Leiden. Scientific Reports, 9(1), 5233. 10.1038/s41598-019-41695-z
- Page, L. et al. (1999). PageRank Citation Ranking. Stanford InfoLab. 10.59350/zdg2m-tv281

24. De Martim, B. (2025). SAT-Graph RAG for Legal Norms.
arXiv:2505.00039. 10.3233/faia251598