Where Does Legal AI Fail? Evaluating RAG Pipelines

Yongjae Kim
Korea Advanced Institute of Science and Technology
The Graduate School of Culture Technology
Daejeon, South Korea
yongjaeya@kaist.ac.kr

Wonjae Lee
Korea Advanced Institute of Science and Technology
The Graduate School of Culture Technology
Daejeon, South Korea
wnjlee@kaist.ac.kr

Abstract

Retrieval Augmented Generation (RAG) frameworks have been integrated into LLMs for QA tasks, which have enabled lower levels of hallucination during the generation stage. Despite ample research in evaluating and detecting hallucination, research regarding its source in the legal domain is still minimal. In this paper, we break down each component of the RAG framework and analyze how they affect the quality of the generation. We use the expert answers of 1,039 real Korean civil complaints provided by the Korean Ministry of Economy and Finance to analyze the quality of generation across 12 experimental configurations that vary in retrieval, reranking, and generation components. To evaluate generation quality, we employ LLM-as-Judge comparing model outputs against expert responses across three dimensions: conclusion consistency, correctness and incorrectness. We validate our findings through logistic regression analysis, establishing statistical relationships between component metrics (Recall@200, NDCG@20) and generation consistency. Despite differing performance across models, two patterns emerge from the generation evaluation: increased retrieval performance correlates with decreased generation quality, and increased reranker performance correlates with higher generation quality. Our work questions the conventional wisdom that "more context is better" and provides a systematic attribution framework for finding the origin of hallucination in legal RAG systems.

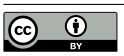
CCS Concepts

• **Information systems** → **Retrieval models and ranking**; • **Computing methodologies** → **Natural language processing**; • **Applied computing** → *Law*.

Keywords

RAG, Hallucination, IR, Legal Chatbots, NLP

ACM Reference Format:
Yongjae Kim and Wonjae Lee. 2025. Where Does Legal AI Fail? Evaluating RAG Pipelines. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761151



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761151

1 Introduction

While Retrieval Augmented Generation frameworks have proven an effective approach in mitigating hallucination across various domains [11], understanding which specific components cause failures remains underexplored. This limitation becomes critical in specialized domains like legal AI, where recent studies reveal 60-80% hallucination rates [3, 14], yet systematic attribution of these failures to specific RAG components has not been established.

The legal domain presents particularly challenging conditions for AI systems, characterized by complex jargon, jurisdictional nuance, and interpretive frameworks where approximations are unacceptable [3, 14]. Real-world legal Q&A requires complex reasoning with specific legal citations, and we separately evaluate retriever and reranker contributions rather than treating retrieval as monolithic.

We focus on Korean legal data, which offers unique advantages for RAG component analysis: the civil complaint system provides natural pairings of colloquial citizen queries with expert legal responses, and Korea's codified legal system with hierarchical statute structures makes component failures more visible as systems must connect general principles with specific regulations scattered throughout the legal code.

This research addresses three specific questions: (1) Which RAG components contribute most to hallucination in legal domain applications? (2) How do retrieval coverage and document ranking quality differentially impact generation accuracy? (3) Do conventional assumptions about retrieval optimization hold for complex legal reasoning tasks? Through controlled component variation across 12 experimental configurations using Korean legal expert responses as ground truth, we isolate hallucination sources and challenge the conventional "more context is better" paradigm.

This research reveals counterintuitive component interactions specific to legal applications, particularly our finding that increased retrieval performance can paradoxically decrease generation quality in this domain. This challenges conventional assumptions that "more context is better" in information retrieval systems [7, 13]. By systematically comparing AI-generated outputs with human expert responses for civil complaints, this study establishes an empirically grounded methodological approach that enables precise identification of technical limitations in legal AI systems, providing actionable insights for legal technology development while advancing scholarly understanding of RAG component interactions in specialized knowledge domains [21].

2 Related Works

2.1 Hallucination Detection to Attribution

Hallucination research in LLMs has progressed significantly in detection but remains limited in source identification. Maynez et al.

(2020) established key distinctions between factual hallucinations (generating false statements) and faithfulness hallucinations (deviating from instructions) [16]. In this paper, we focus specifically on factual hallucinations. While detection frameworks like FactScore [17] and ensemble approaches [5] achieve strong performance, they treat RAG systems as black boxes without identifying which components cause failures.

This limitation becomes critical in legal contexts. Dahl et al. (2024) classify hallucinations as response inconsistencies with the prompt (mischaracterization of opinions), training corpus (creative argumentation), or facts of the world (misstatement of law) [3]. While the first two types might occasionally benefit lawyers, factual infidelity to the controlling legal landscape is intolerable. Dahl et al. (2024) identify four contributing causes of this critical third type: naive retrieval, inapplicable authority, reasoning error, and sycophancy. Building on this, Magesh et al. (2024) further categorize factual hallucinations into closed-domain (misrepresenting provided documents), open-domain training-based (fabricating legal precedents), and open-domain factual (general false statements) [14]. Recent evaluations reveal 60-80% hallucination rates in legal tasks, with even "hallucination-free" legal AI tools showing one in six hallucinated answers [3, 14]. Despite these findings, research has not systematically analyzed whether failures originate from retrieval, reranking, or generation components.

2.2 RAG Component Attribution

Recent studies have sought to disentangle the contributions of individual RAG components to overall system performance. Glass et al. (2022) showed that cross-encoder rerankers substantially improve accuracy by correcting initial retrieval errors and surfacing more relevant content [6]. However, their analysis was limited to simple factoid questions and did not extend to tasks requiring complex reasoning or justification. Similarly, Cuconasu et al. (2024) demonstrated that introducing controlled retrieval "noise" can paradoxically improve accuracy, suggesting that exhaustive retrieval may be suboptimal [2]. Yet, their methodology treated retrieval as a single black-box component, without disentangling the roles of retrievers and rerankers.

Hsia et al. (2024) reported that reader performance follows a non-monotonic trend with respect to retrieval scope: performance improves initially with more context (signal), but then declines as irrelevant information (noise) accumulates [8]. This effect varies by domain, with specialized domains showing distinct sensitivity thresholds. In legal tasks, Pipitone and Alami (2024) found that minimal snippet retrieval outperforms full-document retrieval, which often exceeds context limits and introduces semantic drift [18]. These findings reinforce the need for domain-specific, component-level analysis of retrieval pipelines.

Despite its critical role of rerankers, their role remains underexamined in component attribution. Recent work shows that hybrid architectures employing sophisticated rerankers yield the greatest reductions in hallucination [15]. Nonetheless, most attribution studies still treat the retrieval stage as monolithic, obscuring the distinct effects of the retriever and reranker on final generation quality.

2.3 Legal Domain Applications

Legal QA tasks demands high precision due to domain complexity, formal language, and strict interpretive standards [14]. Standard evaluation metrics fail to capture factual fidelity in this setting. To address this, Ryu et al. (2023) proposed Eval-RAG, a retrieval-augmented evaluation method that more closely aligns with expert (lawyer) judgment, revealing that traditional metrics significantly underestimate error rates in legal outputs [20].

2.4 Research Gap

Three key limitations persist in the current literature. First, most attribution studies focus on shallow factoid queries, neglecting tasks that require multi-hop reasoning and statutory justification. Second, retrieval pipelines are often treated as monolithic, obscuring the individual roles of retrievers and rerankers. Third, domain-specific behaviors in legal contexts remain underexplored, despite growing evidence that general-purpose RAG strategies fail to transfer reliably. Our work addresses these limitations by performing fine-grained attribution across retriever, reranker, and generator components, using complex legal QA tasks grounded in statutory interpretation. This enables precise localization of failure modes in specialized domains and informs the development of more reliable legal AI systems.

3 Methodology

3.1 Problem Formulation

Given a legal query q and a RAG system $\mathcal{RS}$ comprising retriever R, reranker K, and generator G, we seek to attribute hallucinations to specific pipeline components. Unlike existing detection frameworks that produce binary classifications $h \in \{0, 1\}$ indicating presence or absence of hallucination, our approach identifies the origin $o \in \{R, K, G\}$ of each hallucination instance. Formally, for a query q, the RAG pipeline executes:

$$D_r = R(q, C) \quad \text{(Retrieved documents)} \tag{1}$$

$$D_k = K(q, D_r) \quad \text{(Reranked documents)} \tag{2}$$

$$y = G(q, D_k) \quad \text{(Generated response)} \tag{3}$$

where C represents the legal corpus. The generated response y is evaluated against expert responses y^* that serve as ground truth for each query. Our framework decomposes hallucination attribution into component-specific performance functions f_R, f_K, and f_G, establishing statistical associations relationships between component metrics and response accuracy measured against y^*, where each function $f_i : \mathcal{M}_i \to [0, 1]$ maps component-specific metrics $\mathcal{M}_i$ to accuracy probabilities.

3.2 Component Attribution Framework

Our methodology isolates component contributions through a systematic ablation study with controlled experimentation. We decompose the RAG pipeline into three modular components and evaluate their individual and joint contributions to hallucination.

3.2.1 Component Isolation Protocol: For retrieval isolation, we evaluate three distinct strategies while holding downstream components constant. Dense retrieval employs semantic similarity using cosine similarity, sparse retrieval uses lexical matching through

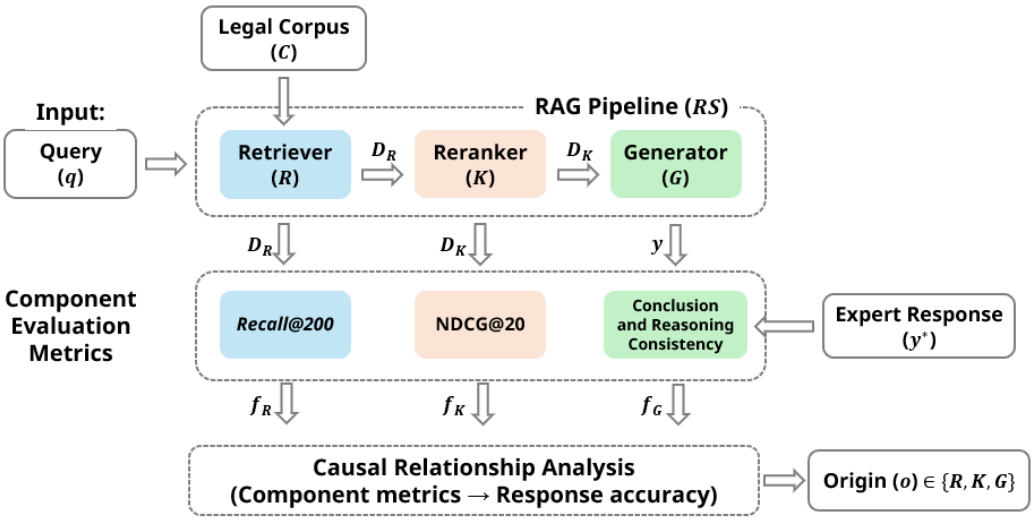


Figure 1: The framework processes a legal query (q) through a three-component pipeline comprising retriever (R), reranker (K), and generator (G), with each component transforming the data flow from legal corpus (C) to final response (y). Our framework attributes hallucinations to specific pipeline components ($o \in R, K, G$) by comparing generated responses against expert ground truth ($y*$). Component-specific performance functions (f_R, f_K, f_G) establish associative relationships between individual component metrics and response accuracy, enabling precise identification of hallucination origins.

BM25 with domain-specific tokenization, and hybrid retrieval combines both approaches using a 70:30 weighted distribution with score normalization. For each strategy, we measure Recall@k for k $\in$ {10, 50, 100, 200} against expert-identified relevant documents, quantifying the retriever's ability to surface critical legal provisions independent of ranking quality.

The reranker analysis compares generation outcomes across two conditions: vanilla reranking using a pre-trained cross-encoder without domain adaptation, and fine-tuned reranking with a cross-encoder adapted on legal query-statute pairs. The reranker processes query-document concatenations through a transformer architecture, outputting relevance scores that determine final document ordering. We measure NDCG@20 to capture both relevance and ranking quality.

For generator evaluation, we process identical top-20 document sets through two models. GPT-4o serves as a closed-source baseline with strong multilingual capabilities, while EXAONE-3.5 represents an open-source alternative specifically optimized for Korean language understanding. Both models receive structured prompts instructing legal reasoning with explicit citation requirements, with generation temperature fixed at 0.1 for reproducibility.

3.2.2 Attribution Analysis: To attribute hallucinations to specific components, we employ a two-stage analysis. In the first stage, we measure performance isolation for each configuration (r, k, g) where $r \in$ {dense, sparse, hybrid}, $k \in$ {vanilla, fine-tuned}, and $g \in$ {GPT-4o, EXAONE-3.5}. We compute retrieval metrics (Recall@200, HitRate@200), ranking metrics (NDCG@20, post-reranking Recall@20), and generation metrics (Correctness, Conclusion Consistency, Incorrectness).

The second stage applies statistical attribution through logistic regression modeling. This quantifies each component's contribution to final accuracy. This factorial design yields 12 experimental configurations, enabling systematic analysis of component interactions. By varying one component while holding others constant, we isolate associative pathways from retrieval through generation to identify precise failure modes in legal RAG systems.

3.3 Evaluation Metrics

3.3.1 Component-Specific Metrics: Retriever evaluation employs *Recall@k* as the primary metric, with *HitRate@k* serving as supplementary reference, evaluated at multiple depths (k $\in$ 10, 50, 100, 200) to understand coverage patterns. This metric was selected because the retriever's fundamental purpose is to identify all relevant legal content. In legal contexts, missing a critical statute can invalidate an entire response, making comprehensive coverage essential. Document relevance is determined by expert-identified legal statutes cited in ground truth responses (detailed in Section 4.2).

Reranker performance assessment utilizes NDCG@20 as the primary metric to capture both relevance and ranking quality. This metric evaluates whether the most important laws are prioritized at the top. In legal applications where context windows are limited, proper prioritization determines which statutes the generative model will primarily reference.

3.3.2 Generation Quality Assessment: Generation evaluation distinguishes between conclusion consistency and comprehensive correctness through formal assessment functions. Let y denote the model output and y^* represent the expert ground truth response.

Conclusion consistency assesses whether equivalent legal determinations are reached:

$$C_{\text{conclusion}}(y, y^*) = \begin{cases} 1 & \text{if legal conclusions match} \\ 0 & \text{otherwise} \end{cases}$$

Correctness requires alignment in both conclusions and underlying legal argumentation:

$$C_{\text{correctness}}(y, y^*) = \begin{cases} 1 & \text{if both conclusion and reasoning match} \\ 0 & \text{otherwise} \end{cases}$$

Incorrectness captures cases where responses definitively fail to align:

$$C_{\text{incorrectness}}(y, y^*) = \begin{cases} 1 & \text{if conclusion and reasoning both fail} \\ 0 & \text{otherwise} \end{cases}$$

These functions map response pairs $(y, y^*) \to \{0, 1\}$, where y represents the model-generated response and y^* represents the expert ground truth response. We employ *GPT-4o* as an automated evaluator following the LLM-as-judge paradigm [22]. While *GPT-4o* serves as both generator and evaluator in different configurations, we mitigate potential bias partially through structured comparison against expert-created ground truth rather than open-ended assessment. To address position bias, each evaluation is performed twice with reversed presentation orders:

$$\text{Final Assessment} = \begin{cases} \text{correct} & \text{if both orders = 1} \\ \text{incorrect} & \text{if both orders = 0} \\ \text{inconclusive} & \text{otherwise} \end{cases}$$

We classify cases as "inconclusive" when judgments differ across presentation orders, "incorrect" when the LLM consistently determines non-alignment across both orders, and "correct" when consistent alignment is found.

3.4 Statistical Analysis Framework

Component attribution employs logistic regression to quantify relationships between pipeline metrics and response accuracy:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \cdot \text{Recall}_{@200} + \beta_2 \cdot \text{NDCG}_{@20} + \varepsilon$$

where $Y \in \{0, 1\}$ represents binary response accuracy with 1 indicating correct and 0 indicating incorrect responses. The intercept term $\beta_0 \in \mathbb{R}$ captures the baseline log-odds of accuracy, while the regression coefficients $\beta_1, \beta_2 \in \mathbb{R}$ quantify the effect of each metric on the log-odds of accuracy. The predictor variables $\text{Recall}_{@200} \in [0, 1]$ and $\text{NDCG}_{@20} \in [0, 1]$ are normalized metric values, and ε represents the error term assumed to follow a standard logistic distribution. The coefficients β_1 and β_2 indicate the change in log-odds of generating a correct response for each unit increase in the respective metric, where positive coefficients suggest that improvements in the metric increase generation accuracy while negative coefficients indicate an inverse relationship.

Multicollinearity diagnostics confirm variance inflation factors below 5.0 across all configurations, validating the statistical independence of predictor variables. Model estimation uses maximum likelihood optimization with BFGS algorithms for each experimental configuration.

4 Experiments

4.1 Experimental Setup

Our experimental methodology employs a systematic factorial design with 12 configurations combining three retrieval strategies (dense, sparse, and hybrid), two reranker variants (vanilla and fine-tuned), and two generation models (*GPT-4o* and *EXAONE-3.5*). We evaluate each configuration across the RAG pipeline: retriever performance, reranker effectiveness, and generation accuracy.

4.2 Dataset Construction

4.2.1 Legal Corpus Compilation: We constructed a Korean legal corpus using the open data API[1] provided by the Korean Ministry of Government Legislation. The corpus encompasses 230 statutes (법령) and 23,215 enforcement decrees (시행령) under the jurisdiction of the Ministry of Economy and Finance (기획재정부), covering taxation, public finance, and economic regulations frequently referenced in civil complaints.

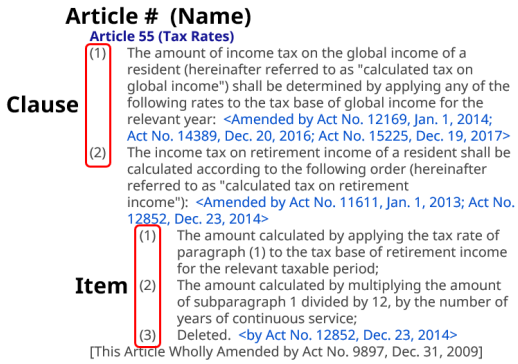


Figure 2: Retrieval Unit Structure Example

Documents are segmented at the article level (조), with subdivisions concatenated to maintain semantic coherence. This yields 41,389 distinct textual chunks. To preserve domain-specific terminology, we inject 3,847 compound legal terms from the Ministry's legal terminology database into the *Kiwi morphological analyzer*[2], ensuring terms like "손해배상책임" (liability for damages) remain unsegmented.

4.2.2 Civil Complaint Dataset Curation: The dataset derives from the e-People platform (국민신문고), which hosts various types of civil complaints[3]. From 252,187 total complaints (2019-2023), we first removed spam through lexical deduplication, leaving 30,212 valid entries. To focus on cases requiring legal interpretation, we filtered for "question complaints" (질의민원) from 2023, yielding 1,306 queries that specifically request legal explanations.

[1] https://open.law.go.kr/LSO/main.do
[2] https://github.com/bab2min/Kiwi
[3] We are preparing a synthetic, privacy-preserving release of the dataset and replication artifacts.

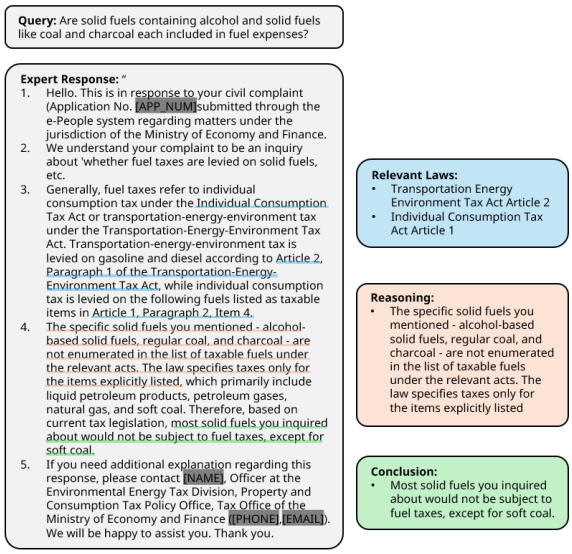


Figure 3: Civil Complaint Structure Example

Final preprocessing excluded responses shorter than 200 characters as potentially lacking substantive legal content, queries explicitly mentioning law names to avoid biasing retrieval evaluation, and queries referencing non-textual content such as images or attachments. This resulted in 1,039 complaint-response pairs, each containing a citizen's colloquial query paired with an expert government response with legal citations. All complaints undergo anonymization with personally identifiable information replaced by standardized placeholders as seen in Figure 3. We employ a 70-30 train-test split for reranker fine-tuning validation.

4.3 Component Implementation

4.3.1 Retrieval Module Architecture: Dense retrieval employs the *BGE-M3 multilingual embedding model*[4] [1, 12], generating 8,192-dimensional vector representations. We utilize *FAISS* (Facebook AI Similarity Search[5]) [4, 9] for efficient vector storage and retrieval due to its optimized indexing structures that enable sub-linear search time even with millions of vectors. Standard cosine similarity between query and document vectors identifies conceptually related content despite terminological variations between colloquial queries and formal legal language.

Sparse retrieval utilizes the BM25 probabilistic ranking algorithm applied to Kiwi-tokenized representations, with hyperparameters k_1 (Term Frequency Saturation Parameter) = 1.2 and b (Length Normalization Parameter) = 0.75 optimized for legal text. Legal dictionary augmentation ensures that domain-specific compounds receive appropriate weighting during lexical matching.

The hybrid approach combines both methodologies through a weighted aggregation scheme. We retrieve 200 candidates from each method, then merge results using a 70:30 distribution between dense and sparse candidates. Score normalization employs logarithmic transformation to reconcile the disparate scaling between cosine similarity and BM25 scores.[6]

4.3.2 Reranking Module Design: The reranking component employs the *BGE-reranker-v2-m3*[7] cross-encoder architecture based on *XLM-RoBERTa* foundations. Unlike bi-encoder approaches that independently encode queries and documents, the cross-encoder processes concatenated query-document pairs to directly model relevance.

For domain adaptation, we fine-tune the base reranker on Korean legal query-statute pairs derived from our complaint dataset. Training employs contrastive loss optimization:

$$\mathcal{L} = -\log \frac{\exp(s(q, d^+)/\tau)}{\exp(s(q, d^+)/\tau) + \sum_{j=1}^{N} \exp(s(q, d_j^-)/\tau)}$$

where $s(q, d) \in \mathbb{R}$ represents the relevance score between query q and document d computed by the cross-encoder. The positive documents d^+ correspond to expert-cited legal provisions from the ground truth responses, while negative samples d_j^- are drawn from the retrieved set that were not cited by experts. The summation includes N negative samples per positive example, and the temperature parameter $\tau = 0.05$ controls the sharpness of the score distribution.

During inference, the reranker evaluates each retrieved document by concatenating the law reference with its content before computing relevance scores, with the top-k documents forwarded to the generation module.

4.3.3 Generation Module Configuration: We evaluate *GPT-4o* (OpenAI) as the closed-source baseline, representing state-of-the-art multilingual capabilities, and *EXAONE-3.5-Instruct*[8] (LG AI) as the open-source alternative specifically optimized for Korean language understanding [19]. This comparison examines whether domain-specific optimization in open models can compensate for the resource advantages of closed systems in specialized applications.

Both models receive structured prompts containing the user query and top-20 reranked documents, with prompts instructing models to assume a legal assistant role and generate Korean responses with formal tone and explicit statutory citations. *GPT-4o* receives prompts in English, while *EXAONE-3.5* receives Korean-translated versions to optimize for its language-specific training.

4.3.4 Implementation Details: The pipeline was implemented using the LangChain framework for RAG orchestration, with *FAISS* for vector storage and HuggingFace Transformers for model inference. For reproducibility, we employed consistent random seeds across all experiments and maintained temperature settings at 0.1 for deterministic generation. *EXAONE-3.5* was loaded using 4-bit quantization to fit within available GPU memory while maintaining inference quality.

[4]https://huggingface.co/BAAI/bge-m3
[5]https://github.com/facebookresearch/faiss

[6]Note that end-to-end hybrid retrievers like *ColBERT* [10] were deliberately excluded from our evaluation. While these methods integrate retrieval and reranking into a single architecture, this integration would prevent the component isolation necessary for our attribution analysis.
[7]https://huggingface.co/BAAI/bge-reranker-v2-m3
[8]https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-32B-Instruct

5 Results

5.1 Component Performance in Isolation

5.1.1 Retrieval Strategy Effectiveness: Table 1 presents the comparative performance of retrieval strategies. The hybrid retrieval approach achieved marginally better performance at all levels of k, though differences remain modest across strategies.

Table 1: Comparative Performance of Retrieval Strategies

Retriever	k=50		k=100		k=200	
	Recall	HitRate	Recall	HitRate	Recall	HitRate
Dense	0.2300	0.3662	0.2705	0.4247	0.3168	0.4756
Sparse	0.2136	0.3380	0.2501	0.3879	0.2906	0.4377
Hybrid	**0.2435**	**0.3824**	**0.2939**	**0.4442**	**0.3176**	**0.4810**

Critically, all retrieval strategies failed to surface relevant provisions for approximately 52-56% of queries within the top 200 results. This systematic retrieval gap stems from the fundamental reliance of dense, sparse, and hybrid approaches on lexical and semantic similarity. In legal contexts, citizen queries often describe situational problems without using technical legal terminology, creating a semantic gap with formal statutory language. This fundamental mismatch between colloquial problem descriptions and technical legal provisions creates an inherent vulnerability for hallucination, regardless of downstream processing sophistication.

5.1.2 Reranking Impact: The reranker module substantially enhanced retrieval effectiveness, with hybrid retrieval showing the most dramatic improvement (12% increase in Recall@20).

Table 2: Reranker Performance Metrics

Configuration	NDCG@20	Recall@20
dense (w/o reranker)	–	0.1871
dense_finetuned	**0.2291**	0.2605
dense_vanilla	0.2160	**0.2663**
sparse (w/o reranker)	–	0.1672
sparse_finetuned	**0.2109**	**0.2479**
sparse_vanilla	0.1383	0.2045
hybrid (w/o reranker)	–	0.1685
hybrid_finetuned	**0.2247**	**0.2898**
hybrid_vanilla	0.2137	0.2881

Table 2 demonstrates that fine-tuned rerankers generally achieved superior NDCG@20 scores across retrieval strategies, indicating better prioritization of legally relevant documents.

5.2 Generation Performance and Hallucination Patterns

Table 3 presents generation performance across all experimental configurations for *GPT-4o* and *EXAONE-3.5*. Baseline zero-shot performance establishes the foundation for comparison, with *GPT-4o* achieving 10.36% correctness and *EXAONE-3.5* achieving 9.50%. RAG implementation demonstrates modest correctness improvements across both models. More importantly, RAG configurations dramatically reduce incorrect response rates, with *GPT-4o* dropping from 48.50% to 41.21-43.44% and *EXAONE-3.5* showing reduction from 58.61% to 24.18-28.04%.

Table 3: Generation Performance (%)

Config	Correct	Incorrect	Inconcl.	Same Concl.
GPT-4o				
dense_ft.	13.81	42.67	43.51	30.62
dense_van.	12.51	41.21	46.28	30.78
hybrid_ft.	13.20	42.13	44.67	29.85
hybrid_van.	12.66	42.67	44.67	29.32
sparse_ft.	13.35	43.44	43.21	30.01
sparse_van.	13.81	42.44	43.75	28.93
zero-shot	10.36	48.50	41.14	26.78
EXAONE-3.5				
dense_ft.	11.28	27.24	61.47	17.73
dense_van.	11.59	27.40	61.01	18.19
hybrid_ft.	14.20	24.36	61.44	14.43
hybrid_van.	11.94	28.04	60.02	17.10
sparse_ft.	12.30	26.90	60.80	16.68
sparse_van.	14.40	24.18	61.42	14.71
zero-shot	9.50	58.61	31.89	20.62

ft. = Finetuned, van. = Vanilla, Inconcl. = Inconclusive

While *GPT-4o* demonstrated higher average correctness (13.23%) compared to *EXAONE-3.5* (12.62%), *EXAONE-3.5* exhibits greater performance variability across configurations, achieving both the highest individual correctness and lower baseline performance. *GPT-4o* maintains more consistent but moderate performance across configurations, a crucial characteristic for commercial deployment where predictable performance across diverse use cases outweighs peak performance in specific configurations. This consistency-performance tradeoff reflects the different design priorities between commercial closed models and research-oriented open models.

The models exhibit distinctly different error patterns. *GPT-4o* produces more definitively incorrect responses (41.21-43.44%), while *EXAONE-3.5* shows substantially higher inconclusive rates (60.02-61.47%). This suggests that the closed model generates more confidently wrong answers, while the open model produces more ambiguous responses when uncertain. The same conclusion rates remain relatively evenly distributed across both models.

While fine-tuned rerankers generally outperform vanilla configurations, sparse retrieval presents a notable exception, consistently performing better with vanilla rerankers across both models. This suggests that lexical matching benefits less from domain-specific reranking optimization, potentially because exact term matches already provide strong relevance signals that fine-tuning may inadvertently dilute.

5.3 RAG Component Effects on Generation Quality

Table 4 presents logistic regression results examining relationships between retrieval metrics and generation outcomes across three measures: conclusion consistency, comprehensive correctness, and incorrect responses. The analysis reveals distinct component contributions and challenges conventional retrieval assumptions.

Table 4: Logistic Regression: Retrieval Metrics Predicting Accuracy and Consistency

Model	Configuration	Metric	Conclusion Consistency		Correctness		Incorrectness	
			Coef.	OR [p-value]	Coef.	OR [p-value]	Coef.	OR [p-value]
GPT-4o	dense_finetuned	NDCG@20	**1.09***	**2.97 [0.019]**	**1.64***	**5.16 [0.018]**	**-0.98***	**0.38 [0.027]**
		Recall@200	-0.43	0.65 [0.274]	**-1.30***	**0.27 [0.038]**	0.14	1.15 [0.691]
	dense_vanilla	NDCG@20	**1.10***	**3.00 [0.022]**	0.83	2.30 [0.161]	**-2.01****	**0.13 [0.000]**
		Recall@200	-0.18	0.84 [0.617]	0.15	1.17 [0.732]	0.53	1.69 [0.123]
	hybrid_finetuned	NDCG@20	**1.02***	**2.76 [0.034]**	**1.41***	**4.11 [0.026]**	**-2.00****	**0.14 [0.000]**
		Recall@200	-0.15	0.86 [0.673]	-0.59	0.55 [0.239]	0.61	1.85 [0.073]
	hybrid_vanilla	NDCG@20	**1.26***	**3.54 [0.013]**	1.01	2.74 [0.121]	**-1.67****	**0.19 [0.002]**
		Recall@200	-0.21	0.81 [0.547]	-0.20	0.82 [0.667]	0.36	1.43 [0.286]
	sparse_finetuned	NDCG@20	**1.14***	**3.13 [0.035]**	1.10	3.01 [0.084]	**-1.56****	**0.21 [0.005]**
		Recall@200	-0.17	0.84 [0.683]	0.17	1.18 [0.747]	0.34	1.40 [0.417]
	sparse_vanilla	NDCG@20	0.82	2.28 [0.051]	0.78	2.17 [0.110]	**-1.22****	**0.30 [0.008]**
		Recall@200	0.25	1.28 [0.369]	0.49	1.63 [0.139]	-0.12	0.89 [0.665]
EXAONE-3.5	dense_finetuned	NDCG@20	**1.15***	**3.16 [0.028]**	0.89	2.45 [0.150]	**-2.59****	**0.08 [0.000]**
		Recall@200	-0.10	0.90 [0.823]	-0.08	0.92 [0.881]	**0.91***	**2.50 [0.013]**
	dense_vanilla	NDCG@20	0.68	1.97 [0.218]	**1.66***	**5.25 [0.014]**	**-1.52****	**0.22 [0.002]**
		Recall@200	0.00	1.00 [0.999]	-0.73	0.48 [0.180]	0.46	1.58 [0.172]
	hybrid_finetuned	NDCG@20	-1.06	0.35 [0.110]	-0.92	0.40 [0.169]	0.07	1.08 [0.877]
		Recall@200	0.70	2.01 [0.113]	0.59	1.81 [0.186]	0.33	1.39 [0.334]
	hybrid_vanilla	NDCG@20	**1.19***	**3.29 [0.048]**	0.85	2.35 [0.201]	-0.78	0.46 [0.121]
		Recall@200	-0.42	0.66 [0.334]	-0.03	0.97 [0.942]	0.18	1.19 [0.589]
	sparse_finetuned	NDCG@20	0.67	1.96 [0.310]	0.53	1.69 [0.467]	-0.47	0.62 [0.368]
		Recall@200	-0.20	0.82 [0.707]	-0.00	1.00 [0.994]	0.02	1.02 [0.962]
	sparse_vanilla	NDCG@20	0.85	2.33 [0.216]	0.95	2.59 [0.169]	-0.04	0.96 [0.924]
		Recall@200	**-1.05***	**0.35 [0.031]**	**-1.09***	**0.33 [0.027]**	0.42	1.53 [0.116]

Note: * **p < 0.05,** ** **p < 0.01; Coef = coefficient; OR = odds ratio**

5.3.1 Component Effect Size Analysis: Reranking performance exhibits strong, statistically significant effects across most *GPT-4o* configurations. More specifically, in the dense_finetuned configuration, a unit increase in NDCG@20 more than quintuples the odds of generating a correct answer (OR = 5.16, p = .018), while cutting incorrect outputs by over 60% (OR = 0.38, p = .027). This pattern holds consistently across configurations, with NDCG@20 showing the most significant effects Notably, reranking effects are strongest in dense retrieval configurations across both models, suggesting that semantic retrieval benefits most from sophisticated reranking. Conversely, sparse retrieval shows the most variable reranker sensitivity, with vanilla approaches sometimes outperforming fine-tuned versions, indicating that lexical matching may conflict with learned reranking patterns. In contrast, higher retrieval coverage shows a consistent directional trend of degrading correctness or increasing hallucination, though with weaker statistical support. This is particularly evident in the dense_finetuned configuration of *GPT-4o*, where higher recall reduces the odds of correctness by 73% (OR = 0.27, p = .038). This establishes a strong and consistent association between reranking quality and legal RAG performance, suggesting that reranking is the dominant component influencing system reliability within our experimental framework.

5.3.2 The Retrieval-Generation Paradox: The statistical analysis reveals a paradox that challenges conventional assumptions about retrievers. Rather than following the traditional "more context is better" paradigm, higher retrieval coverage can actively degrade generation quality in legal reasoning tasks. This aligns with Hsia et al. (2024), who found reader performance exhibits a non-monotonic relationship with retrieval scope, where performance initially improves with more contexts due to increased "signal" but eventually degrades as "noise" overwhelms relevant information. Similarly, findings by Cuconasu et al. (2024) and Pipitone and Alami (2024), who demonstrated benefits of controlled noise or minimal snippets are aligned with our results indicating an inherent tradeoff where excessive document retrieval overwhelms the generation process.

The consistently negative Recall@200 coefficients across multiple configurations, combined with strong positive NDCG@20 effects, demonstrate that retrieval coverage and document ranking quality have opposing relationships with generation accuracy. Our results extend these findings by empirically proving that effective generation depends not on retrieving all relevant documents, but on retrieving the most important documents in optimal order. This suggests RAG systems should prioritize sophisticated ranking mechanisms over exhaustive retrieval strategies.

5.3.3 Model-Specific Response Patterns: While this paradox affects both models, their responses to component quality differ substantially. The two language models exhibit markedly different sensitivity patterns to component quality, revealing that component contributions vary significantly across model architectures. *GPT-4o* demonstrates consistent reranker sensitivity across configurations, with NDCG@20 showing significant positive effects across configurations for conclusion consistency and strong effects for reducing incorrect responses. The model exhibits predictable component interactions, making it more suitable for systematic optimization.

EXAONE-3.5 exhibits configuration-dependent behavior, with reranker effects varying significantly across retrieval strategies. Dense and hybrid vanilla configurations show strong NDCG@20 effects (OR 3.16-3.29), while sparse configurations show minimal sensitivity. However, *EXAONE-3.5* demonstrates the strongest negative effects from excessive retrieval in sparse_vanilla configuration (OR 0.33-0.35), suggesting greater sensitivity to retrieval-generation imbalance in specialized domain applications. This reveals that component contributions vary significantly across model architectures, with GPT-4o showing more predictable optimization patterns than the domain-specialized open-source alternative.

5.4 Hallucination Attribution Analysis

To understand failure mechanisms underlying incorrect responses, we categorize error sources using Dahl et al. (2024) and Magesh et al. (2024)'s hallucination taxonomy. For responses marked as incorrect by our evaluation framework, we identify whether failures stem from incorrect interpretation of retrieved authority, incorrect retrieval of relevant legal provisions, or sycophantic responses prioritizing agreeableness over accuracy.

Table 5: Distribution of Hallucination Causes

Configuration	Incorrect Interpretation	Incorrect Retrieval	Sycophancy
GPT-4o			
dense_finetuned	72.59%	26.82%	0.59%
dense_vanilla	72.10%	27.22%	0.68%
hybrid_finetuned	73.00%	26.50%	0.50%
hybrid_vanilla	73.29%	25.95%	0.76%
sparse_finetuned	71.81%	27.54%	0.66%
sparse_vanilla	71.48%	27.84%	0.69%
EXAONE-3.5			
dense_finetuned	65.45%	34.14%	0.40%
dense_vanilla	66.31%	33.29%	0.40%
hybrid_finetuned	39.89%	59.93%	0.18%
hybrid_vanilla	62.06%	37.53%	0.41%
sparse_finetuned	65.46%	34.26%	0.28%
sparse_vanilla	40.34%	59.49%	0.17%

5.4.1 Model-Specific Error Patterns: GPT-4o demonstrates consistent error patterns across all configurations, with interpretation failures dominating retrieval failures by approximately 3:1 ratios. This suggests the model successfully incorporates retrieved documents but struggles with precise legal application. Sycophancy remains negligible across both models, indicating appropriate objectivity in legal domain responses.

EXAONE-3.5 exhibits configuration-dependent error patterns. While most configurations show interpretation-dominated failures similar to *GPT-4o*, hybrid fine-tuned and sparse vanilla configurations flip to retrieval-dominated failures. This bifurcation indicates that *EXAONE-3.5*'s context utilization capabilities vary significantly with retrieval pipeline design.

5.4.2 Component Attribution Implications: These patterns directly address which components contribute most to system failures. For *GPT-4o*, the consistent interpretation dominance suggests that improving retrieval quality alone will not address the primary failure source. For *EXAONE-3.5*, failure attribution depends critically on configuration choice, with certain retrieval-reranker combinations leading to fundamental context utilization breakdowns rather than reasoning errors.

6 Limitations

While our study provides systematic insights into RAG component failures in legal applications, several limitations should guide interpretation of our findings.

Our factorial design provides quasi-experimental evidence but cannot establish definitive causality. Despite controlling for query-specific confounds through identical test sets and temporal precedence in the pipeline architecture, we cannot eliminate all potential confounding variables such as component-document interactions. We interpret our regression coefficients as robust associations rather than causal effects. Additionally, multiple comparisons across 12 configurations could increase Type I error risk. However, the consistent directional patternsparticularly opposing NDCG@20 (positive) and Recall@200 (negative) effectssuggest true relationships rather than random false positives.

Methodologically, using GPT-4o as both generator and evaluator could introduce systematic bias despite our mitigation strategies. The high inconclusive rates (43-61%) suggest binary correctness assessments may oversimplify legal reasoning nuances. Our exclusion of end-to-end architectures like ColBERT, while necessary for component isolation, prevents comparison with state-of-the-art unified systems. Human legal expert validation would strengthen these findings.

7 Conclusion and Future Directions

This paper conducted component-level analysis of RAG pipelines in Korean legal applications using expert-annotated civil complaints. Our experimental design across multiple retriever, reranker, and generator configurations revealed how individual components influence system performance and hallucination patterns.

Our analysis challenges conventional information retrieval assumptions by demonstrating that increased retrieval coverage can actively degrade generation quality due to information overload. Conversely, reranker effectiveness consistently correlates with improved outcomes, establishing document ranking quality as the primary driver of system performance. Statistical analysis revealed distinct model-specific sensitivities: *GPT-4o* exhibits predictable component interactions across configurations, while *EXAONE-3.5* shows configuration-dependent behavior significantly influenced by specific retrieval-reranker combinations.

Hallucination attribution analysis identified different failure mechanisms across models. *GPT-4o* predominantly faces interpretive challenges despite adequate retrieval quality, while *EXAONE-3.5*'s errors frequently stem from retrieval inadequacies in specific configurations. These patterns enable targeted optimization strategies rather than generic system improvements.

Building on the retrieval-generation paradox, future work should investigate optimal retrieval scope thresholds for different legal reasoning tasks. The component attribution methodology developed here should be extended to other specialized domains to validate generalizability. Additionally, research should explore hybrid evaluation frameworks that balance automated assessment with domain expert validation to address the high inconclusive rates observed in legal evaluation tasks.

This study establishes a systematic framework for diagnosing component-specific failures in domain-specialized RAG systems, providing actionable insights for developing more reliable legal AI applications.

Acknowledgements

This research was supported by the Ministry of Science and ICT, Korea, under the Graduate School of Metaverse Convergence support program supervised by the Institute of Information & Communications Technology Planning & Evaluation under Grant No. IITP-2022(2025)-RS-2022-00156435.

A Generation Prompt Template

The following prompt template was used for both models:

```
You are an expert Korean government civil service
assistant, specializing in answering citizen inquiries
about regulations, laws, and governmental procedures.
You are known for your careful legal reasoning and
attention to detail when analyzing legal questions.
When answering the following inquiry, please follow
this structured reasoning approach:
1) First, carefully analyze the query to identify the
   precise legal issues being raised
2) Identify which specific laws, regulations, or
   statutes are most relevant to these issues
3) Examine how these legal provisions apply to the
   particular situation described in the query
4) Consider multiple interpretations of the law if
   the situation is ambiguous
5) Formulate a clear answer based on the most
   authoritative legal sources
6) Include specific legal citations to support your
   conclusions
Provide responses that are concise, focused on
legal/regulatory issues, based strictly on relevant
laws and regulations, and written in formal,
respectful Korean. Avoid personal opinions or
interpretations unless specifically addressing gaps
or ambiguities in the law.
Query: {query}
Relevant Legal Context: {retrieved_legal_documents}
Based on the legal context provided above, please
carefully analyze the query and provide a well-
reasoned answer. In your response:
1. Identify the specific legal provisions most
   relevant to the inquiry
2. Explain how these provisions apply to the
   situation described
3. If there are multiple interpretations possible,
   acknowledge this and explain which is most
   applicable
4. Support your conclusions with specific citations
   to the legal texts provided
5. If the context doesn't contain enough information
   for a complete answer, explain what additional
   legal information would be needed and why
Write your response in Korean using formal legal
language appropriate for government communications.
Present your analysis as a single cohesive paragraph
that demonstrates clear legal reasoning. Answer in Korean:
```

B LLM-as-Judge Evaluation Prompt

The prompt included bias mitigation through dual evaluation:

```
You are an expert legal evaluator. Compare two
responses independently to a legal question
objectively and determine if they contain the same
conclusion and reasoning.
Question: {question}
{first_label}: {first_response}
{second_label}: {second_response}
Evaluate these responses by analyzing their substance
thoroughly. First identify the main conclusions and
key reasoning points in each response. Avoid any
position biases and ensure that the order in which
the responses were presented does not influence your
decision. Do not allow the length of the responses
to influence your evaluation. Do not favor certain
names of the assistants. Be as objective as possible.
Then provide your evaluation:
####Same Conclusion####
[Answer only in Yes or No]
####Same Reasoning####
[Answer only in Yes or No]
####Reason for Error####
[If errors exist, choose from: "Incorrect Retrieval",
"Incorrect interpretation of authority", "Sycophancy".
If multiple apply, separate with commas. If none
apply, write "None"]
Guidelines for error types:
- Incorrect Retrieval: The response cites or relies
  on wrong, irrelevant, or non-existent laws or
  regulations
- Incorrect interpretation of authority: The response
  identifies relevant legal authority but misinterprets
  or misapplies it
- Sycophancy: The response appears to agree with or
  flatter the user rather than provide accurate legal
  analysis
```

References

[1] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216 [cs.CL]

[2] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ACM, 719729. https://doi.org/10.1145/3626772.3657834

[3] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (Jan. 2024), 64–93. https://doi.org/10.1093/jla/laae003 arXiv:2401.01301 [cs]

[4] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazar, Maria Lomeli, Lucas Hosseini, and Herv Jgou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[5] Grant C. Forbes, Parth Katlana, and Zeydy Ortiz. 2023. Metric Ensembles For Hallucination Detection. arXiv:2310.10495 [cs.CL] https://arxiv.org/abs/2310.10495

[6] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, Rerank, Generate. arXiv:2207.06300 [cs.CL] https://arxiv.org/abs/2207.06300

[7] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv:2002.08909 [cs.CL] https://arxiv.org/abs/2002.08909

[8] Jennifer Hsia, Afreen Shaikh, Zhiruo Wang, and Graham Neubig. 2024. RAGGED: Towards Informed Design of Retrieval Augmented Generation Systems. arXiv:2403.09040 [cs.CL] https://arxiv.org/abs/2403.09040

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[10] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. arXiv:2004.12832 [cs.IR] https://arxiv.org/abs/2004.12832

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kttler, Mike Lewis, Wen tau Yih, Tim Rocktschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[12] Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making Large Language Models A Better Foundation For Dense Retrieval. arXiv:2312.15503 [cs.CL]

[13] Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. arXiv:2403.10446 [cs.CL] https://arxiv.org/abs/2403.10446

[14] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2024. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. https://doi.org/10.48550/arXiv.2405.20362 arXiv:2405.20362 [cs]

[15] Chandana Sree Mala, Gizem Gezici, and Fosca Giannotti. 2025. Hybrid Retrieval for Hallucination Mitigation in Large Language Models: A Comparative Analysis. arXiv:2504.05324 [cs.IR] https://arxiv.org/abs/2504.05324

[16] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. arXiv:2005.00661 [cs.CL] https://arxiv.org/abs/2005.00661

[17] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv:2305.14251 [cs.CL] https://arxiv.org/abs/2305.14251

[18] Nicholas Pipitone and Ghita Houir Alami. 2024. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. arXiv:2408.10343 [cs.AI] https://arxiv.org/abs/2408.10343

[19] LG AI Research. 2024. EXAONE 3.5: Series of Large Language Models for Real-world Use Cases. *arXiv preprint arXiv:https://arxiv.org/abs/2412.04862* (2024).

[20] Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, Daniel Preoiuc-Pietro, Catalina Goanta, Ilias Chalkidis, Leslie Barrett, Gerasimos Spanakis, and Nikolaos Aletras (Eds.). Association for Computational Linguistics, Singapore, 132–137. https://doi.org/10.18653/v1/2023.nllp-1.13

[21] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2022. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. arXiv:2210.02627 [cs.CL] https://arxiv.org/abs/2210.02627

[22] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685

C Generative AI Usage Disclosure

This research involved limited use of generative AI tools in specific, well-defined tasks as detailed below. All core research contributions, methodology design, data analysis, and substantive writing were conducted without AI assistance.

Literature Discovery: ChatGPT Deep Research was used during the initial research phase to identify existing work in retrieval-augmented generation component analysis. This process led to the discovery of Cuconasu et al. (2024)'s work on retrieval noise effects, which informed our research direction. All subsequent literature review and analysis were conducted manually without AI assistance.

Data Format Conversion: Generative AI was used to convert CSV tables containing experimental results into LaTeX table format for inclusion in the manuscript. The underlying data, analysis, and interpretation remained entirely human-generated.

Data Preprocessing Support: During the dataset curation phase, generative AI was employed at a high level to identify obvious formatting inconsistencies and data quality issues in the Korean civil complaint dataset. However, AI could not detect all preprocessing requirements, and comprehensive data cleaning, validation, and preparation were performed manually.

Text Reduction: Due to conference length limitations, generative AI was used to reduce the Conclusion section to approximately 80% of its original length. Following this process, substantial manual editing was performed to ensure retention of the authors' writing style, technical accuracy, and key research contributions.

All research design, experimental methodology, statistical analysis, results interpretation, and primary writing were conducted entirely by the authors without generative AI assistance.