From Legal Texts to Normative Intent: Bridging Intent Gaps for Superior Legal Article Retrieval

Anonymous ACL submission

Abstract

Superior legal article retrieval is essential for legislative drafting, as it helps drafters identify higher-level articles that a new regulation must follow, thereby reducing conflicts and improving cross-level consistency. Accurate matching in superior legal article retrieval remains challenging, largely because current models have little prior exposure to superior legal articles, making their query reformulations deviate from real legal texts. To address this, we propose a normative-intent-based retrieval framework for superior legal article retrieval. Instead of directly generating or matching articles, our method generates the core normative intent of each article and incorporates it as an intermediate signal in the retrieval process, reformulating superior legal article retrieval as intent-level alignment rather than surface-level textual matching. This shift transforms the reliance on domain-specific knowledge of superior legal articles into a demand for intent generation, a capability that large language models are better suited to generalize across domains. We conduct experiments on SLARD, a Chinese superior legal article retrieval dataset. Results show that incorporating normative intent improves average Recall@K by 5.64 percentage points and MRR@5 by 6.73 points compared with the text-only dense retriever. This study advances superior legal article retrieval and supports legal text understanding in legislative digitalization using technologies based on large language models.

1 Introduction

As society evolves, new regulations are continuously established to adapt to social and economic changes [Donelan, 2022]. During the legislative drafting process, retrieving relevant superior legal articles from existing laws serves as an essential foundation. Enacted by higher-level legislative bodies such as national or provincial authorities, these articles provide guiding principles for lower-level regulations and help ensure legal consistency. This hierarchical structure plays a critical role in preventing normative conflicts and maintaining coherence within the legal system [Vinx, 2007]. Accordingly, superior legal article retrieval is an important component of legislative review and legal interpretation, supporting consistency across regulatory frameworks.

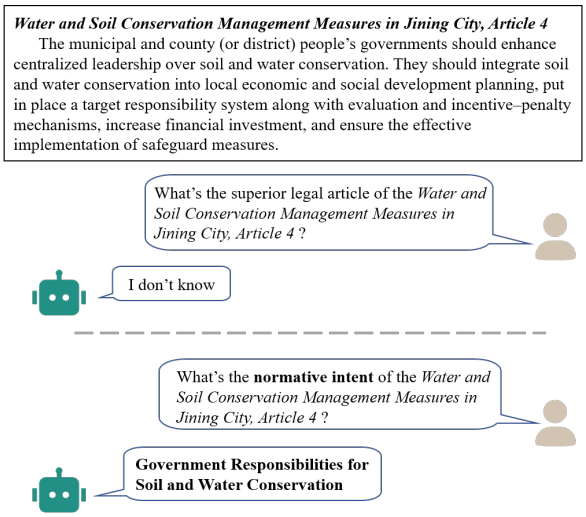

Figure 1: Illustrative Example of Normative Intent for Superior Legal Article Retrieval. LLMs often lack prior knowledge of the correct superior legal basis for a municipal article, but can reliably infer its normative intent. Intent extraction thus provides a stable semantic bridge for cross-level retrieval, reducing noise from level-specific textual details.

Recent studies in digital legislation have explored natural language processing (NLP) and information retrieval techniques to support superior legal article retrieval [Sansone and Sperlí, 2022; Opmane et al., 2019]. Early approaches typically rely on lexical matching, rule-based methods, or feature-driven models. While these methods provide a foundation for automated legal retrieval, they often struggle to capture high-level legal intent and semantic relations across articles. More

recent work incorporates Large Language Models (LLMs) to enhance semantic understanding. However, due to the lack of explicit prior knowledge about superior legal articles, LLM-based methods may produce generated representations that deviate from the normative expressions of legal articles. Such deviations can lead to inaccurate matching, especially in scenarios requiring precise alignment with legally authoritative texts.

To address the above challenges, we propose a Normative Intent-Based Retrieval framework for superior legal article retrieval, which introduces the concept of normative intent into LLM-based superior legal article retrieval and effectively improves retrieval accuracy. Specifically, we introduce normative intent as an intermediate representation that abstracts each superior legal article into a concise expression of its normative intent and integrates this representation into the retrieval process. As illustrated in Figure 1, instead of requiring LLMs to generate or match legally precise articles, the proposed method reformulates superior legal article retrieval as an intent-level alignment problem mediated by normative intent. This reformulation fundamentally shifts the role of LLMs from producing domain-specific legal text to summarizing and aligning high-level normative intent, thereby reducing the dependence on specialized legal knowledge. By leveraging normative intent as a semantic bridge between queries and superior legal articles, the proposed approach enables more stable and accurate retrieval under the hierarchical and normative constraints of legal texts.

We conduct experiments on SLARD, a Chinese superior legal article retrieval dataset, and evaluate sparse, dense, and LLM-based retrieval baselines using Recall@K and MRR. Results show that incorporating normative intent improves average Recall@K by 5.64 percentage points and MRR@5 by 6.73 points over the text-only dense retriever. These results demonstrate that normative intent provides an effective intermediate alignment signal for bridging municipal-level articles and superior legal articles, thereby improving cross-level legal retrieval. Code and resources are available in an anonymized repository.[1]

The main contributions are summarized as follows:

- We introduce normative intent abstraction for superior legal article retrieval, using LLMs to distill legal articles into concise representations of regulatory purpose and obligation structure.

- We design a normative intent-based retrieval framework that combines intent-level alignment with original article content through weighted fusion.

- We perform extensive experiments on SLARD and provide systematic analyses, confirming the effectiveness and robustness of our approach.

2 Related Work

2.1 Legal Information Retrieval

Legal information retrieval has been studied from several perspectives, including knowledge graph-based methods, graph neural network (GNN)-based methods, and pretrained language model (PLM)-based methods. Knowledge graph-based approaches improve retrieval by explicitly modeling relations among legal provisions, cases, and factual elements [Li et al., 2024; Shukla et al., 2025; Kondo et al., 2025]. However, they often depend on manual or predefined graph construction and are less effective in capturing implicit normative intent across legal hierarchies. GNN-based methods further learn latent relational patterns over legal graphs and have shown effectiveness in modeling case relations and legal logic [Tang et al., 2024; Donabauer and Kruschwitz, 2025], but they mainly focus on local structural interactions rather than cross-level normative alignment.

PLM-based methods enhance legal semantic representation through domain adaptation, continual pretraining, and knowledge-enhanced objectives [Li et al., 2023; Niyogi and Bhattacharya, 2024; Nigam et al., 2025; Ghosh et al., 2024]. Although these methods improve surface-level semantic matching, they still lack explicit modeling of normative intent and hierarchical legal relationships, which are crucial for superior legal article retrieval. In contrast, our work introduces normative intent as an intermediate abstraction, enabling retrieval models to align legal articles according to their regulatory purpose and obligation structure rather than relying only on structural links or textual similarity.

[1] https://anonymous.4open.science/r/paper-artifacts-E6F8/

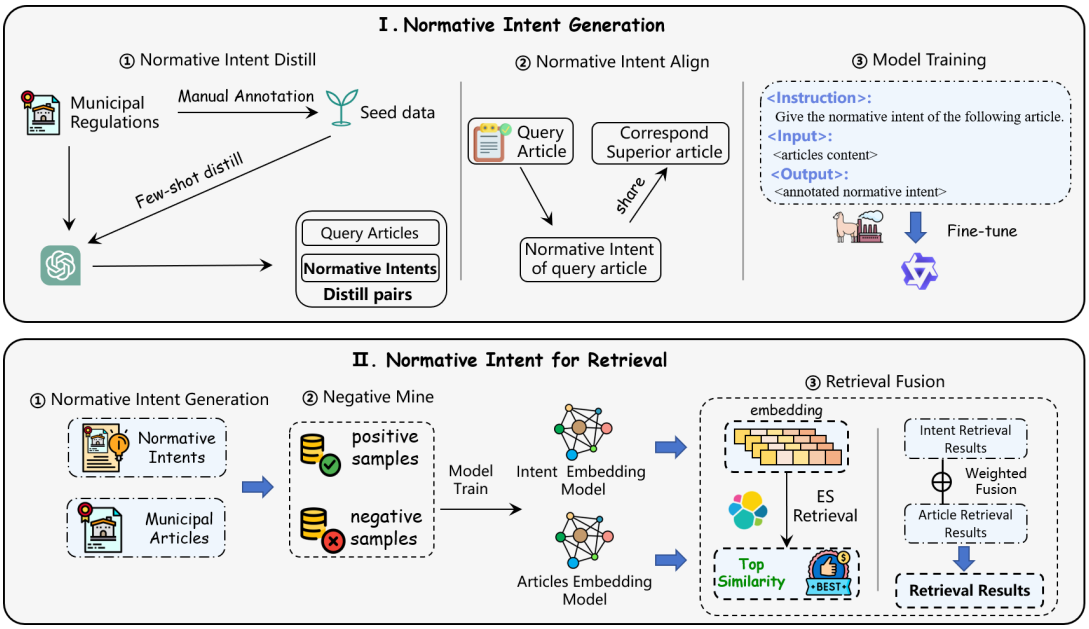

Figure 2: Overall Framework of Superior Legal Article Retrieval Based on Normative Intent

2.2 Large Language Models for Information Retrieval

LLMs have recently been applied to information retrieval through dense representation learning, query augmentation, and generative retrieval. Vector-based methods leverage LLMs to produce dense representations for queries and documents through zero-shot prompting, lightweight adaptation, or contrastive training [Jiang et al., 2024; Lei et al., 2024; Liu et al., 2024; Zhang et al., 2025]. Although these methods improve general semantic matching, superior legal article retrieval requires alignment beyond textual relatedness, including regulatory purpose, obligation structure, and hierarchical legal authority. Existing LLM-based dense retrieval methods usually do not explicitly encode these legal-specific factors, and may still be affected by level-specific drafting details.

Another line of work uses LLMs for query augmentation or generative retrieval, where expanded queries, pseudo-documents, or document identifiers are generated to improve retrieval coverage [Gao et al., 2023; Wang et al., 2023; Li et al., 2025]. However, LLMs may lack prior knowledge of real upper-level legal provisions, causing generated content to contain structural inconsistency, inaccurate citations, or incomplete normative logic. Such outputs may fail to correspond to authoritative superior legal texts. Our method differs from these approaches by avoiding direct generation of superior legal articles. Instead, it uses LLMs to produce normative intent, a compact intent-level representation that serves as a semantic bridge between municipal-level articles and superior legal articles.

3 Proposed Method

3.1 Method Overview

We propose a normative intent-based retrieval framework for superior legal article retrieval across hierarchical legal systems. As shown in Figure 2, the framework contains two components. First, normative intent generation abstracts each legal article into its regulatory purpose and obligation structure, reducing the influence of wording, granularity, and drafting conventions. Second, normative intent-based retrieval uses the generated intent as an intermediate alignment signal and combines it with original article content for ranking. In this way, superior legal article retrieval is reformulated as an intent-level alignment problem rather than purely surface-level textual matching.

3.2 Normative Intent Generation

This section describes how we construct normative intent representations for legal articles. Let

$$\mathcal{A} = \{a_1, a_2, \ldots, a_N\}$$

denote the set of municipal-level legal articles, where each a_i corresponds to an individual legal article.

Normative Intent Distillation. We first manually annotate a small subset of municipal-level

legal articles with high-quality normative intents, forming a seed dataset:

$$\mathcal{D}_{\text{seed}} = \{(a_i, \hat{n}_i)\}_{i=1}^{M}, \quad M \ll N,$$

where $\hat{n}_i$ denotes the manually annotated normative intent of article a_i. Based on this seed set, we employ a strong LLM to perform few-shot normative intent distillation. Specifically, examples from $\mathcal{D}_{\text{seed}}$ are provided as in-context demonstrations to guide the model in generating normative intent for unlabeled articles. For an unlabeled article a_j, its normative intent is generated as

$$n_j = f_\theta(a_j \mid \mathcal{D}_{\text{seed}}),$$

where f_θ denotes a parameterized normative intent generation model and n_j is the generated intent. This process yields an expanded article–intent dataset:

$$\mathcal{D}_{\text{distill}} = \{(a_j, n_j)\},$$

enabling scalable construction of normative intent annotations.

Normative Intent Alignment. To align query articles with their corresponding superior legal articles at the intent level, we introduce a normative intent consistency assumption. Let q denote a query article and

$$\mathcal{S}^+(q) = \{s_1, s_2, \ldots, s_K\}$$

denote its ground-truth superior legal articles. Due to the inherent consistency of legal norms across hierarchical levels, we assume that a query article and its true superior articles share the same normative intent. Accordingly, the normative intent of each superior article s_k is set as

$$n_{s_k} = n_q, \quad \forall s_k \in \mathcal{S}^+(q),$$

where n_q denotes the normative intent of the query article. For superior articles that are not aligned with any query article, their normative intents are obtained from the aforementioned distillation process. This alignment strategy constructs a unified intent space spanning municipal-level and superior-level legal articles. This alignment is used only to construct supervision signals on the training split. During evaluation, no query–superior relevance labels are used; the normative intents of both test queries and candidate superior articles are generated solely from their own texts.

Model Training. After normative intent distillation and alignment, we obtain the final training dataset:

$$\mathcal{D}_{\text{NI}} = \{(a, n_a)\},$$

where n_a denotes the normative intent associated with article a. We fine-tune open-source LLMs to learn the mapping from legal articles to normative intents. The model is trained in an autoregressive generation manner using the standard cross-entropy loss:

$$\mathcal{L}_{\text{NI}} = -\sum_{(a,n_a)\in\mathcal{D}_{\text{NI}}} \sum_{t=1}^{|n_a|} \log P_\theta(n_{a,t} \mid a, n_{a,<t}),$$

where $n_{a,t}$ denotes the t-th token in the normative intent sequence.

3.3 Normative Intent–Based Retrieval

This section describes the superior legal article retrieval module based on normative intent. Instead of using a single fused representation, we adopt a dual-path retrieval framework that performs retrieval on normative intent and original article content in parallel, and merges the results with late score fusion. This design enables the model to jointly leverage abstract normative alignment and fine-grained textual distinctions.

Let q denote a query article and

$$\mathcal{S} = \{s_1, s_2, \ldots, s_L\}$$

denote the set of candidate superior legal articles.

Normative Intent Generation. Using the trained normative intent generation model introduced in Section 3.2, we generate normative intents for all query articles and candidate superior legal articles. For any article a, its normative intent is produced as

$$n_a = f_\theta(a),$$

where f_θ denotes the trained normative intent generation model and n_a is the resulting intent representation. This process provides a unified intent-level view for both query articles and superior articles, forming the basis for intent-based retrieval.

Hard Negative Mining. To improve the discriminative ability of the retrieval models, we incorporate a hard negative mining strategy during training. For a query article q, let $\mathcal{S}^+(q) \subset \mathcal{S}$ denote its ground-truth superior legal articles, while the remaining candidates are treated as potential negatives. Based on the current retrieval models, we

perform preliminary retrieval using both normative intent representations and article text representations. Highly ranked candidates that do not belong to $\mathcal{S}^+(q)$ are selected as hard negatives, forming the negative set $\mathcal{S}^-(q)$. By contrasting true superior articles with these challenging negatives, the retrieval models are encouraged to distinguish normatively similar but legally irrelevant articles.

Retrieval Fusion. At inference time, the intent-based path computes $\mathrm{sim}_I(q, s)$ between n_q and n_s, while the article-based path computes $\mathrm{sim}_T(q, s)$ over original article texts.

Since normative intent emphasizes alignment at the regulatory-function level, while article text captures fine-grained legal distinctions, we combine the two retrieval signals through weighted score fusion. The final retrieval score is defined as

$$r(q, s) = \alpha \ \mathrm{sim}_I(q, s) + (1 - \alpha) \ \mathrm{sim}_T(q, s),$$

where $\alpha \in [0, 1]$ is a hyperparameter that controls the relative contribution of normative intent-based retrieval and original article content-based retrieval. Candidates are ranked according to the fused scores, and the top-K superior legal articles are returned as the final retrieval results.

4 Experiments

4.1 Dataset

We conduct experiments on SLARD, the first publicly available dataset specifically designed for superior legal article retrieval [Chen et al., 2025]. SLARD contains 2,627 municipal-level legal articles as queries and 9,184 candidate superior legal articles, including 2,976 provincial-level and 6,208 national-level articles, covering 32 legal topic categories. Since SLARD is annotated at the provision level, it requires models to capture fine-grained hierarchical legal relationships and reflects practical scenarios in legislative drafting and legality review.

4.2 Evaluation Metrics

We evaluate retrieval performance using Recall@K and Mean Reciprocal Rank (MRR). Recall@K measures the proportion of relevant superior legal articles retrieved within the top-K results:

$$\mathrm{Recall@K} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{|\mathcal{R}_q \cap \mathcal{S}_q^K|}{|\mathcal{R}_q|}, \qquad (1)$$

where $\mathcal{Q}$ is the query set, $\mathcal{R}_q$ denotes the relevant superior legal articles for query q, and $\mathcal{S}_q^K$ denotes the top-K retrieved results.

MRR evaluates ranking quality by considering the rank of the first relevant superior legal article within the top five results:

$$\mathrm{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{\mathrm{rank}_q}, \qquad (2)$$

where rank_q is the rank position of the first relevant result for query q. If no relevant article appears in the top five, the reciprocal rank is set to zero.

4.3 Baselines

We compare our method with representative retrieval baselines from four categories. Sparse retrieval models include BM25 [Robertson and Zaragoza, 2009] and DocT5Query [Nogueira and Lin, 2019]. Generic pre-trained retrieval models include Chinese-BERT-WWM, Chinese-RoBERTa-WWM [Cui et al., 2021], and Lawformer [Xiao et al., 2021]. Retrieval-oriented pre-trained models include DPR [Karpukhin et al., 2020], RetroMAE [Xiao et al., 2022], and ColBERT [Khattab and Zaharia, 2020]. LLM-based retrieval models include HyDE [Gao et al., 2023] and Query2Doc [Wang et al., 2023]. These baselines cover lexical matching, dense retrieval, legal-domain pretraining, late interaction, and LLM-based query expansion.

4.4 Training Details

All experiments are conducted on 4 NVIDIA RTX 3090 GPUs. We split SLARD into training and testing sets with a ratio of 7:3. To avoid label leakage, intent alignment based on ground-truth municipal–superior article pairs is applied only to the training split; during testing, normative intents are generated solely from article texts without relevance labels.

For normative intent generation, we fine-tune Qwen2.5-7B, Qwen3-8B, GLM4-9B, DeepSeek-7B, DeepSeek-R1-0528-Qwen3-8B, and Ministral-8B-Instruct-2410 with LoRA, and evaluate GPT-5-mini and Claude Sonnet 4.5 in a prompting-only setting. Unless otherwise specified, GLM4-9B is used as the default generator in sensitivity analyses. For retrieval, we use bge-base-zh-v1.5 as the embedding backbone and train it with contrastive learning and hard negatives. During inference, the final score is obtained by fusing full-text similarity and normative-intent similarity.

Method	R@1	R@3	R@5	R@10	R@15	R@20	MRR@5
BM25	44.62	70.17	76.65	83.13	85.94	88.14	57.69
DocT5Query	38.14	60.88	67.60	75.92	79.22	81.30	49.94
Chinese-BERT-WWM	25.55	33.01	33.99	37.04	39.24	39.85	29.12
Chinese-RoBERTa-WWM	27.63	35.82	38.51	41.69	43.15	43.64	31.94
Lawformer	26.16	34.60	37.41	40.34	41.56	42.42	30.57
DPR	47.19	74.33	81.30	87.90	90.10	91.81	61.07
RetroMAE	47.19	74.57	81.66	88.75	90.59	92.30	61.18
ColBERT	40.22	66.63	73.47	80.81	84.47	86.55	53.76
HyDE	11.74	20.54	24.08	30.56	36.19	39.85	16.58
Query2Doc (Sparse)	26.28	45.35	53.67	63.81	69.80	72.00	36.57
Query2Doc (Dense)	29.95	52.08	61.37	72.25	77.02	78.97	41.96
Qwen2.5-7B	49.02	80.68	87.16	92.54	94.62	**95.48**	65.04
Qwen3-8B	50.49	81.54	88.14	**93.03**	**94.99**	95.35	66.32
GLM4-9B	51.47	**82.52**	88.02	92.79	94.62	95.23	66.94
DeepSeek-7B	50.12	76.89	84.23	90.22	92.18	93.52	64.04
DeepSeek-R1[2]	50.49	79.22	84.60	88.75	90.22	92.05	64.83
Ministral-8B	49.76	78.61	83.37	89.61	91.44	92.79	64.13
GPT-5-mini	51.59	79.58	85.45	89.36	91.81	93.03	65.72
Claude Sonnet 4.5	50.12	76.77	82.89	89.00	91.20	93.15	63.87
Ours	**52.57**	**82.52**	**88.88**	92.91	94.50	95.11	**67.80**

Table 1: Main retrieval results on the SLARD dataset. Best results are bolded; second-best are underlined.

5 Results and Discussion

5.1 Main Results

Table 1 reports the main retrieval results on the SLARD dataset. The upper part compares sparse, dense, and LLM-based retrieval baselines, the middle part reports variants that concatenate original article content with normative intents generated by different LLMs, and the last row presents our weighted fusion method. Ours achieves the strongest top-rank retrieval performance, with the highest R@1 of 52.57 and MRR@5 of 67.80, while maintaining a competitive R@10 of 92.91. Compared with DPR, it improves R@1 from 47.19 to 52.57 and MRR@5 from 61.07 to 67.80, indicating clear gains in both top-rank accuracy and ranking quality. It also outperforms RetroMAE, a strong retrieval-oriented baseline, by 6.62 points in MRR@5.

These improvements show that normative intent serves as an effective intermediate alignment signal for superior legal article retrieval. By abstracting legal provisions into their regulatory purpose and obligation structure, normative intent reduces the influence of level-specific drafting details in full-text representations. This allows the retriever to better align municipal-level articles with their corresponding superior legal articles. In contrast, HyDE and Query2Doc perform substantially worse than retrieval-based baselines, with MRR@5 scores below 42. This suggests that directly generating hypothetical documents or expanded queries is unreliable for this task, where relevance depends on hierarchical legal consistency rather than free-form semantic relatedness.

The middle part of Table 1 further shows that the benefit of normative intent is not tied to a single generator. All concatenation variants improve over the text-only dense retriever in MRR@5. For example, GLM4-9B improves MRR@5 from 61.07 to 66.94 and R@1 from 47.19 to 51.47, while Qwen3-8B achieves strong recall performance with an R@10 of 93.03. Even prompting-only proprietary models, such as GPT-5-mini and Claude Sonnet 4.5, improve MRR@5 to 65.72 and 63.87, respectively. Compared with the strongest concatenation variant, Ours further improves MRR@5 from 66.94 to 67.80 and R@1 from 51.47 to 52.57, confirming that weighted fusion better balances original article content and normative intent than simple concatenation.

[2]DeepSeek-R1 denotes DeepSeek-R1-0528-Qwen3-8B.

Setting	R@1	R@3	R@5	R@10	R@15	R@20	MRR@5
$n = 0$	47.19	74.33	81.30	87.90	90.10	91.81	61.07
$n = 1$	**51.47**	**82.52**	**88.02**	**92.79**	**94.62**	**95.23**	**66.94**
$n = 2$	50.86	77.63	83.37	88.51	91.20	92.42	64.44
$n = 5$	51.34	78.85	83.50	89.12	90.71	91.69	64.99
$n = 10$	48.41	78.24	83.74	88.26	90.22	91.81	63.43
$n = 15$	48.17	76.65	83.13	87.53	89.73	91.81	62.81

Table 2: Sensitivity to intent repetition.

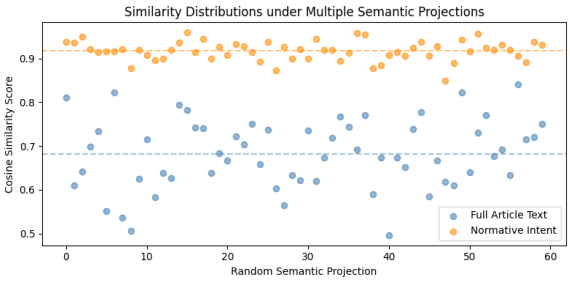

Figure 3: Similarity distributions of full-text and normative intent representations.

Model	Setting	BLEU	ROUGE-L	BERTScore-F1
Qwen2.5-7B	w/o NI	20.04	28.03	78.55
	w/ NI	31.37 (+11.33)	51.27 (+23.24)	79.59 (+1.04)
Qwen3-8B	w/o NI	19.58	20.52	73.55
	w/ NI	31.59 (+12.01)	52.70 (+32.18)	80.11 (+6.56)
GLM4-9B	w/o NI	21.35	28.81	78.66
	w/ NI	33.91 (+12.56)	53.10 (+24.29)	80.05 (+1.39)
Ministral-8B	w/o NI	2.54	16.40	71.12
	w/ NI	31.71 (+29.17)	50.78 (+34.38)	78.70 (+7.58)
DeepSeek-7B	w/o NI	13.94	21.25	75.70
	w/ NI	32.56 (+18.62)	51.96 (+30.71)	79.31 (+3.61)
DeepSeek-R1	w/o NI	19.09	32.98	79.51
	w/ NI	31.54 (+12.45)	52.57 (+19.59)	80.10 (+0.59)
GPT-5	w/o NI	24.65	34.48	75.89
	w/ NI	25.04 (+0.39)	44.07 (+9.59)	81.24 (+5.35)
Claude Sonnet 4.5	w/o NI	19.70	30.03	76.41
	w/ NI	21.49 (+1.79)	45.66 (+15.63)	79.89 (+3.48)

Table 3: Generation-based similarity metrics with and without normative intent.

5.2 Sensitivity Analysis

We further analyze the sensitivity of the model to the amount of normative intent information in the input. Specifically, we concatenate the normative intent with the original article text and repeat the intent n times, where $n \in \{1, 2, 5, 10, 15\}$. This experiment examines whether repeatedly emphasizing normative intent can further improve retrieval performance. As shown in Table 2, adding a single normative intent instance ($n = 1$) substantially improves retrieval over the text-only setting ($n = 0$) across all metrics. However, further repetition does not lead to additional gains and even degrades performance in most cases. This suggests that a concise normative intent is sufficient to provide the key regulatory-purpose signal, while excessive repetition may introduce redundancy and reduce the contribution of fine-grained textual information. This observation further supports our dual-path fusion design, which balances intent-level alignment with article-based discrimination rather than simply increasing the frequency of normative intent tokens.

5.3 Similarity Analysis of Normative Intent Representations

To better understand why normative intent improves retrieval, we examine whether it provides a stable semantic bridge between municipal-level legal articles and their corresponding superior legal articles. We conduct a *multi-view semantic similarity analysis* over aligned municipal–superior article pairs. Given two representations $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we project them into multiple low-dimensional semantic subspaces and compute their cosine similarity in each view:

$$s_i = \cos(\mathbf{R}_i\mathbf{x}, \mathbf{R}_i\mathbf{y}), \tag{3}$$

where $\mathbf{R}_i \in \mathbb{R}^{k\times d}$ is a random projection matrix and $k \ll d$. The resulting distribution $\{s_i\}$ reflects the stability of semantic alignment under different projected views.

As shown in Figure 3, normative-intent representations yield consistently higher similarities and lower variance than full-text representations. This indicates that normative intent provides a more stable alignment signal for municipal–superior legal article pairs. In contrast, full legal texts contain diverse drafting details, such as procedural requirements, implementation conditions, and administrative constraints, which may differ substantially across legislative levels even when the articles regulate the same issue. This makes full-text similarity more sensitive to drafting-level variation.

We further use generation-based similarity metrics to examine whether normative intent reduces the expression gap between municipal-level articles and superior legal articles. BLEU measures n-gram overlap, ROUGE-L captures longest-common-subsequence similarity, and BERTScore-F1 evaluates contextual semantic similarity based on token embeddings. As shown in Table 3, the w/ NI setting consistently improves over w/o NI across all evaluated models and metrics. The gains are especially pronounced on ROUGE-L, where GLM4-9B improves from 28.81 to 53.10, Qwen3-8B from 20.52 to 52.70, and DeepSeek-7B from 21.25 to 51.96. BERTScore-F1 also improves con-

α	R@1	R@3	R@5	R@10	R@15	R@20	MRR@5
0.1	49.88	78.73	84.47	89.98	91.93	93.40	64.41
0.3	50.37	80.07	86.19	90.95	92.67	93.77	65.35
0.5	51.47	81.54	87.53	92.05	93.64	94.62	66.73
0.7	**52.57**	**82.52**	**88.88**	92.91	94.50	95.11	**67.80**
0.9	52.32	**82.52**	**88.88**	**93.15**	**94.87**	**95.60**	67.68

Table 4: Impact of the weighting coefficient α on retrieval performance.

Setting	R@1	R@3	R@5	R@10	MRR@5
Full Method	**52.57**	**82.52**	**88.88**	**92.91**	**67.80**
w/o Normative Intent	47.19	74.33	81.30	87.90	61.07
w/o Article Content	34.35	54.89	62.84	71.03	45.39
w/o Weighted Fusion	51.47	**82.52**	88.02	92.79	66.94

Table 5: Ablation analysis of the proposed method.

sistently, such as from 73.55 to 80.11 for Qwen3-8B and from 75.89 to 81.24 for GPT-5, indicating that the effect is not limited to lexical overlap but extends to contextual semantic alignment. Together with Figure 3, these results show that normative intent provides a compact and hierarchy-agnostic representation of legal meaning, serving as a stable intermediate alignment signal for superior legal article retrieval.

5.4 Impact of the Weighting Coefficient

Table 4 reports the impact of the weighting coefficient α, where a larger α assigns more weight to normative intent-based retrieval. As α increases from 0.1 to 0.7, R@1 improves from 49.88 to 52.57 and MRR@5 improves from 64.41 to 67.80, indicating that normative intent provides an effective alignment signal for ranking relevant superior legal articles at top positions. The best R@1 and MRR@5 are achieved at $\alpha = 0.7$, suggesting that top-rank accuracy benefits from emphasizing intent-level alignment while still retaining sufficient article-based evidence.

When α further increases to 0.9, the model obtains the best R@10, R@15, and R@20, but its MRR@5 slightly decreases. This shows that assigning a very large weight to normative intent can improve broader retrieval coverage, but may weaken fine-grained discrimination among highly similar superior legal articles. Therefore, normative intent and original article content play complementary roles: normative intent captures high-level regulatory purpose and obligation structure, while original article content preserves detailed legal conditions and procedural distinctions. Weighted fusion enables the model to balance these two signals, leading to more accurate and reliable superior legal article retrieval.

5.5 Ablation Analysis

To further verify the effectiveness of normative intent, we conduct an ablation analysis in Table 5. The full method uses both original article content and normative intent, and combines the two retrieval signals through weighted fusion. We compare it with three variants that remove or replace key components.

Removing normative intent leads to clear drops across all metrics, with MRR@5 decreasing from 67.80 to 61.07 and R@10 from 92.91 to 87.90. This confirms that normative intent provides an effective cross-level alignment signal beyond original article content, which is often affected by drafting style, granularity, and level-specific expressions.

The w/o Article Content variant suffers a much larger degradation, with MRR@5 falling to 45.39. This indicates that normative intent cannot replace original article content, because fine-grained legal conditions, procedural requirements, and institutional scopes are still necessary for distinguishing similar superior legal articles. Finally, w/o Weighted Fusion also underperforms the full method, suggesting that weighted fusion is more effective than simple concatenation in balancing intent-level alignment and article-based discrimination. Overall, the results demonstrate that normative intent and original article content are complementary signals for superior legal article retrieval.

6 Conclusion

We propose a retrieval framework that explicitly incorporates normative intent for superior legal article retrieval. By abstracting legal provisions into regulatory purpose and obligation structure, normative intent serves as an intermediate alignment signal between municipal-level articles and superior legal articles. The proposed weighted fusion combines intent-level alignment with full-text signals, balancing high-level normative consistency with fine-grained legal distinctions. Experiments on S-LARD show that our method outperforms strong baselines, while further analyses confirm that normative intent provides a stable and complementary signal for hierarchical legal retrieval. Future work will explore more structured intent representations, hierarchy-aware ranking, and adaptive fusion mechanisms for more complex legal systems.

Limitations

Superior legal article retrieval is closely tied to the structure, terminology, and drafting conventions of specific legal systems. Although our framework is designed around general principles of hierarchical legal alignment, its adaptation to other jurisdictions, languages, or legal domains may require further validation and domain-specific adjustment. Second, the quality of normative intent depends on the intent generation model. Incomplete or overly abstract normative intent may weaken intent-level alignment and affect retrieval performance. Third, normative intent captures regulatory purpose and obligation structure at a high level, but it may omit fine-grained legal conditions, procedural requirements, or institutional scopes that are important for distinguishing highly similar superior legal articles. Therefore, our method uses normative intent as a complementary signal to original article content rather than as a standalone representation. Finally, the proposed system is intended to assist superior legal article retrieval and should not replace expert legal verification.

AI Generation Statement

We used generative AI tools to assist with English proofreading and minor text polishing. All technical content, experimental design, implementation, and results were produced by the authors. The authors reviewed and verified the final manuscript and take full responsibility for its content.

Ethics Statement

The dataset used in this study consists of publicly accessible legal and regulatory documents and contains no personal or sensitive information. The proposed system is intended as an assistive retrieval tool and should not be treated as legal advice or an automated legal decision.

References

Zhe Chen, Pengjie Ren, Fuhui Sun, Xiaoyan Wang, Yujun Li, Siwen Zhao, and Tengyi Yang. 2025. S-LARD: A Chinese superior legal article retrieval dataset. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 740–754.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Gregor Donabauer and Udo Kruschwitz. 2025. A reproducibility study of graph-based legal case retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3135–3144.

Edward Donelan. 2022. *Regulatory Governance: Policy Making, Legislative Drafting and Law Reform*. Springer Nature.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 1762–1777.

Sudipto Ghosh, Devanshu Verma, Balaji Ganesan, Purnima Bindal, Vikas Kumar, and Vasudha Bhatnagar. 2024. InLegalLLaMA: Indian legal knowledge enhanced large language model. In *Proceedings of LKM@IJCAI*, pages 37–46.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Ryoma Kondo, Riona Matsuoka, Takahiro Yoshida, Kazuyuki Yamasawa, and Ryohei Hisano. 2025. Capturing legal reasoning paths from facts to law in court judgments using knowledge graphs. In *Proceedings of the 13th Knowledge Capture Conference 2025*, pages 103–110.

Yibin Lei, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates. 2024. Meta-task prompting elicits embeddings from large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 10141–10157.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. SAILER: Structure-aware pre-trained language model for legal case retrieval. In *Proceedings*

of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1035–1044.

Jun Li, Lu Qian, Peifeng Liu, and Taoxiong Liu. 2024. Construction of legal knowledge graph based on knowledge-enhanced large language models. *Information*, 15(11):666.

Rui Li, Quanyu Dai, Zeyu Zhang, Xu Chen, Zhenhua Dong, and Ji-Rong Wen. 2025. KnowTrace: Bootstrapping iterative retrieval-augmented generation with structured knowledge tracing. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Volume 2*, pages 1470–1480.

Zheng Liu, Chaofan Li, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024. Llama2Vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 3490–3500.

Shubham Kumar Nigam, Deepak Patnaik Balaramamahanthi, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. NYAYAANUMANA and INLEGALLLAMA: The largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11135–11160.

Mitodru Niyogi and Arnab Bhattacharya. 2024. PARAMANU-AYN: Pretrain from scratch or continual pretraining of LLMs for legal domain adaptation? *arXiv preprint arXiv:2403.13681*.

Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery. An MS MARCO Passage Retrieval Task Publication.

Inara Opmane, Juris Balodis, and Rihards Balodis. 2019. Governance of legislative requirements for the development of natural language processing tools. In *MIC 2019: Managing Geostrategic Issues; Proceedings of the Joint International Conference, Opatija, Croatia, 29 May–1 June 2019*, pages 13–27. University of Primorska Press.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Carlo Sansone and Giancarlo Sperlí. 2022. Legal Information Retrieval Systems: State-of-the-art and open issues. *Information Systems*, 106:101967.

Siddharth Shukla, Tanuj Tyagi, Abhay Singh Bisht, Ashish Sharma, and Basant Agarwal. 2025. NyayGraph: A knowledge graph enhanced approach for legal statute identification in Indian law using large language models. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 147–156.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2024. CaseGNN: Graph neural networks for legal case retrieval with text-attributed graphs. In *Proceedings of the European Conference on Information Retrieval*, pages 80–95. Springer.

Lars Vinx. 2007. *Hans Kelsen's Pure Theory of Law: Legality and Legitimacy*. Oxford University Press.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2Doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.