MCOMPASSRAG: Topic Metadata as a Semantic Compass for Paragraph-Level Retrieval

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) systems depend critically on how documents are chunked and searched. Fine-grained chunks can improve retrieval precision but expand the search space, increasing latency and cost; larger chunks reduce the number of candidates but make dense similarity less reliable, as each representation mixes multiple topics and introduces more semantic noise. This trade-off becomes especially limiting in deep research tasks, where retrieval must be both fast and precise across large, heterogeneous corpora. We introduce **MCOMPASSRAG**, a metadata-guided retrieval framework that uses topic-level signals as a semantic compass for selecting relevant evidence. Instead of relying only on cosine similarity between queries and noisy chunk embeddings, MCOMPASSRAG enriches chunk representations with topic metadata in the same embedding space and trains a lightweight retriever through LLM-teacher distillation. At inference time, MCOMPASSRAG performs topic-aware retrieval without additional LLM calls, improving both efficiency and evidence quality. Across six complex retrieval benchmarks, MCOMPASSRAG improves information efficiency by **8.24%** on average with over **5**× **lower latency** than the strongest efficient RAG baselines[1].

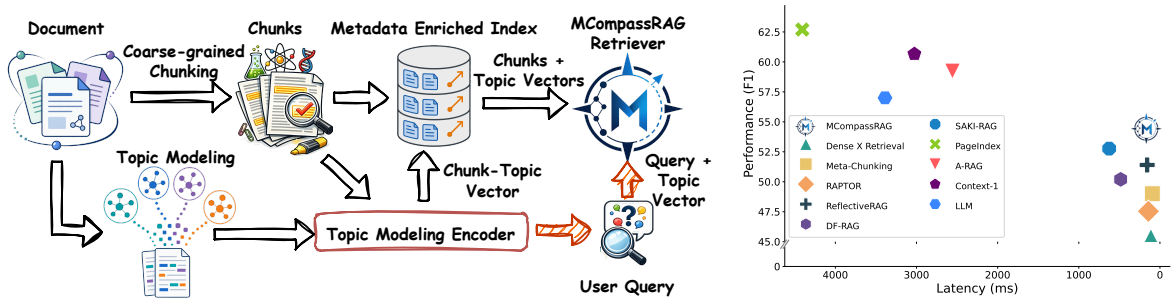
1 Introduction

Retrieval-augmented generation (RAG) has become a standard paradigm for grounding large language models (LLMs) in external knowledge (Lewis et al., 2020; Karpukhin et al., 2020). Yet the efficiency and quality of RAG hinge on a simple but consequential design choice: how documents are divided into retrievable units. This choice becomes especially important in deep research tasks (Zhang et al., 2025b), where systems must search large corpora and often issue many retrieval calls before producing a final answer. Standard dense retrieval over fixed-size chunks (Zhao et al., 2024) faces a granularity trade-off. Fine-grained chunks, such as sentences or short paragraphs, offer precise evidence but greatly increase the number of candidates to index and search. Larger chunks reduce the search space and improve retrieval efficiency, but they mix multiple topics and discourse roles into a single embedding. As a result, similarity scores become noisy: relevant evidence can be diluted by unrelated text, while partially relevant chunks may be retrieved despite containing mostly irrelevant content.

Prior work addresses chunk granularity by either making chunks smaller, more structured, or hierarchically organized. Proposition-level retrieval decomposes documents into atomic units (Chen et al., 2024), LLM-guided segmentation improves chunk boundaries (Zhao et al., 2025b,a), and hierarchical methods such as RAPTOR retrieve across multiple abstraction levels (Sarthi et al., 2024). While effective, these approaches often increase pre-processing cost, require additional indices, or introduce extra scoring and selection stages. LLM-based re-ranking and evidence selection can further improve quality (Tao et al., 2025), but add latency at inference time, which is problematic for deep research agents that repeatedly retrieve evidence over large corpora (Zheng et al., 2025).

In this work, we take a different approach: instead of changing the chunk size dynamically or adding expensive post-retrieval filtering, we make larger chunks more searchable. As shown in Figure 1a, **MCOMPASSRAG** enriches chunk representations with topic metadata that acts as a semantic compass for retrieval. When topic signals are represented in the same embedding space as the retriever, they expose the main semantic directions covered by each chunk, allowing retrieval to look beyond a single noisy cosine similarity

[1] Code, data, and checkpoints will be released publicly upon publication.



(a) (b)

Figure 1: Overview of MCompassRAG. **(a)** MCompassRAG uses coarse chunks for efficiency and enriches them with topic vectors for topic-aware retrieval. At query time, relevant topic information guides retrieval over larger chunks. **(b)** MCompassRAG improves the performance–latency trade-off over strong RAG baselines.

score. At query time, a metadata selection module identifies relevant topic signals, and an abstraction layer converts the selected metadata into a compact query-side representation for topic-aware matching. MCompassRAG is agnostic to the specific topic model, requiring only that topics be embedded in the retriever's semantic space. We train MCompassRAG as an extreme multi-label classifier (Prabhu et al., 2025) using LLM-teacher distillation, where a lightweight student learns to identify multiple relevant chunks from metadata-enriched representations without LLM calls at inference time. This preserves the efficiency advantage of larger chunks while reducing the semantic noise that makes coarse-grained cosine retrieval unreliable. Across six complex retrieval benchmarks, MCompassRAG **improves information efficiency by 8.24%** on average over the strongest non-LLM baseline while running at over **5× lower latency** compared to strong LLM-based RAG baselines, reflecting the efficiency–quality trade-off as shown in Figure 1b.

Our **contributions** are threefold. **First**, we introduce **MCompassRAG**, a metadata-guided retrieval framework that uses selected topic metadata to make coarse-grained retrieval more precise without increasing the search space. **Second**, we propose a metadata selection and abstraction module that identifies query-relevant topical signals from a metadata bank and converts them into a compact retrieval representation. **Third**, we distill an LLM teacher into a lightweight student retriever with an extreme multi-label objective, enabling topic-aware retrieval without inference-time LLM calls.

2 Related Work

Retrieval Granularity and Structured Retrieval in RAG. RAG grounds language model generation in external evidence retrieved before generation (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021). A key design choice is retrieval granularity: fine-grained units improve evidence precision but enlarge the search space and may lose context, while coarse-grained units preserve context and reduce candidates but make dense similarity noisier due to mixed topics and irrelevant content. Prior work addresses this trade-off through alternative retrieval units or index structures, including proposition-level retrieval (Chen et al., 2024), LLM-guided and adaptive chunking (Zhao et al., 2025b,a), query-adaptive granularity selection (Zhang et al., 2026), and hierarchical retrieval across abstraction levels (Sarthi et al., 2024). Other systems enrich retrieved evidence to reduce context fragmentation (Tao et al., 2025) or promote diversity and coverage during selection (Khan et al., 2026). While effective, these methods often require finer-grained indexing, adaptive selection, hierarchical structures, extra scoring stages, or LLM-based filtering. In contrast, MCompassRAG preserves the efficiency of coarse-grained retrieval while making larger chunks more searchable with topic-level metadata.

Semantic Guidance and Efficient Retrieval. A complementary line of work improves RAG by modifying the query or retrieval process rather than the chunking strategy itself. Query augmentation methods such as HyDE (Gao et al., 2023), query expansion (Wang et al., 2023; Zhou et al., 2024), and decomposition-based retrieval (Trivedi et al., 2023; Zheng et al., 2024) aim to better align the query with relevant evidence by generating hypothetical answers, adding related terms, or breaking complex questions into simpler retrieval steps. Adaptive and iterative retrieval methods further refine the evidence set through repeated retrieval, rerank-

ing, or sufficiency checking (Verma et al., 2026). These methods are effective when the query underspecifies the needed evidence, but they often introduce extra inference-time computation. Separately, generation-side efficiency methods compress or reorganize retrieved context after retrieval to reduce decoding cost (Lin et al., 2025; Louis et al., 2026). MCompassRAG is orthogonal to these directions: rather than generating additional query text, repeatedly retrieving, or compressing context after retrieval, it uses corpus-derived topic metadata as a compact semantic guide before retrieval. This guides retrieval toward query-relevant topics without inference-time LLM calls, and remains compatible with query expansion, iterative retrieval, reranking, and context compression.

3 MCompassRAG

As shown in Figure 2, MCompassRAG is a metadata-guided retrieval framework that makes coarse-grained chunks more searchable without increasing the retrieval search space. Given chunks $\mathcal{C} = \{c_1, \ldots, c_N\}$ and a query q, the goal is to retrieve the top-k evidence chunks. Instead of relying only on cosine similarity between query and chunk embeddings, MCompassRAG augments both with topic-level metadata. The framework can use any topic modeling approach whose topics are represented in the retriever's embedding space and that provides chunk-level topic distributions. In our implementation, we adopt CEMTM (Abaskohi et al., 2025), an LLM-distilled topic model that also leverages attention signals to produce document-topic vectors. MCompassRAG first builds a metadata bank by extracting topic vectors for each document chunk. At inference time, it then: **(i)** selects query-relevant metadata from the metadata bank, **(ii)** abstracts the selected metadata into a compact query-side topic representation, and **(iii)** scores query–chunk pairs with a lightweight student classifier.

3.1 Topic Metadata and Metadata Bank

Let $\{\mathbf{t}_k\}_{k=1}^{K}$ denote the topic centroids, where each $\mathbf{t}_k \in \mathbb{R}^d$ lies in the retriever embedding space. Each chunk c is associated with a topic distribution: $\boldsymbol{\theta}_c \in \mathbb{R}^K$, where $\theta_{c,r}$ measures the strength of topic r in chunk c. Since chunks are longer and more informative than queries, their topic distributions can be computed reliably and cached offline. MCompassRAG stores these chunk-level topic distributions in a metadata bank:

$$\mathcal{M} = \{\boldsymbol{\theta}_{c_1}, \ldots, \boldsymbol{\theta}_{c_N}\}. \tag{1}$$

The **metadata bank** represents the topical structure of the corpus and serves as the source of **query-side guidance at inference time**. Given a new query, MCompassRAG does not directly rely on the query's own topic distribution, which may be unreliable due to its short length. Instead, it selects relevant topic distributions from $\mathcal{M}$ and abstracts them into a compact query-side topic representation. This abstraction step reduces bias toward any single selected chunk and produces a smoother topical signal, as described in Section 3.2.

3.2 Metadata Selection and Representation

At inference time, MCompassRAG selects topic metadata from the bank that is relevant to the input query. The query is first encoded by the student encoder, f_ψ:

$$\mathbf{e}_q = f_\psi(q) \in \mathbb{R}^d. \tag{2}$$

We implement the **selection policy** as a lightweight scoring module over the concatenation of the query embedding and each metadata-entry embedding. Each metadata entry $\boldsymbol{\theta}_{c_i}$ is first converted into an embedding-space summary:

$$\mathbf{m}_i = \sum_{k=1}^{K} \theta_{c_i,k} \mathbf{t}_k. \tag{3}$$

The selector assigns an unnormalized compatibility score to each metadata entry based on the topic centroids, t_k:

$$a_i = \mathbf{w}_s^\top [\mathbf{e}_q; \mathbf{m}_i] + b_s, \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation. The scores are converted into a probability distribution over metadata entries using softmax operation:

$$s_i = \frac{\exp(a_i)}{\sum_{j=1}^{N} \exp(a_j)}. \tag{5}$$

The top-L metadata entries according to s_i are selected and passed to the **abstraction module**.

$$\mathbf{H}^{(0)} = [\boldsymbol{\theta}_{\tilde{q}_{j_1}}; \ldots; \boldsymbol{\theta}_{\tilde{q}_{j_L}}] \in \mathbb{R}^{L \times K}. \tag{6}$$

After a two-layer Transformer encoder (Vaswani et al., 2017), the outputs are mean-pooled to form a refined query topic distribution:

$$\hat{\boldsymbol{\theta}}_q = \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{H}_\ell^{(2)}. \tag{7}$$

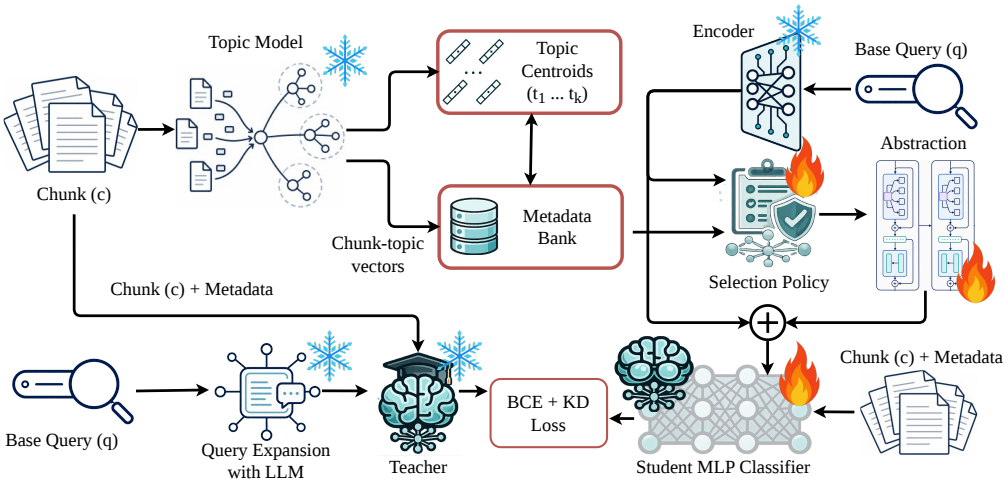


Figure 2: Overview of MCOMPASSRAG. During training, an LLM teacher provides relevance supervision, with query expansion used only as an additional teacher-side metadata signal. The metadata bank is built from chunks, enriched with document-topic vectors and topic centroid embeddings. At inference time, MCOMPASSRAG selects and abstracts query-relevant topic metadata, then scores query–chunk pairs with a lightweight student retriever.

This abstraction step combines complementary topic signals and suppresses redundant or noisy metadata entries and constructs topic-enriched representations for both chunks and queries. For a chunk c, we select the top-L topics from its topic distribution:

$$\mathcal{T}_c = \text{top-M}(\boldsymbol{\theta}_c), \tag{8}$$

and aggregate their topic centroids:

$$\mathbf{g}_c = \sum_{k \in \mathcal{T}_c} \theta_{c,k} \mathbf{t}_k. \tag{9}$$

The final chunk representation is: $\mathbf{r}_c = [\mathbf{e}_c; \mathbf{g}_c]$. Similarly, the refined query topic distribution $\hat{\boldsymbol{\theta}}_q$ is used to build a query-side topic summary with the top-M topics: $\mathbf{r}_q = [\mathbf{e}_q; \mathbf{g}_q]$.

The student retriever scores each query–chunk pair with a three-layer MLP classifier:

$$z(q, c) = \text{MLP}_\phi([\mathbf{r}_q; \mathbf{r}_c]), \tag{10}$$

where $z(q, c)$ is the predicted relevance logit. This formulation casts retrieval as an extreme multi-label classification problem: each chunk is a candidate label, and each query may correspond to multiple relevant chunks.

3.3 Training with LLM-Teacher Distillation

Training data construction. We synthesize training data from the training split of each benchmark. For each dataset, we sample 2,000 chunks and use GPT-4o (OpenAI, 2024) to generate 10 natural queries per chunk, resulting in 20,000 query–chunk pairs before negative sampling. For each sampled chunk c_i, GPT-4o receives the target chunk together with its preceding and following chunks. It first generates a base query q_i whose answer requires evidence from c_i. It then generates an expanded query $\tilde{q}_i$ by adding only background information from the two of the neighboring chunks, without revealing the answer or including answer-specific hints. We use Prompt A.1 for the query expansion.

For relevance supervision, the source chunk is treated as a positive candidate, while negatives are sampled from non-matching chunks. We include both random negatives and hard negatives, where hard negatives are retrieved using `Qwen3-Embedding-4B` (Zhang et al., 2025c) as high-similarity chunks that the LLM teacher judges as not useful for answering the query. GPT-4o is then used as an LLM teacher: given the expanded query $\tilde{q}_i$ and a candidate chunk, it predicts whether the chunk provides direct or supporting evidence for answering the query (see Prompt A.2). The resulting hard label $y \in \{0, 1\}$ and teacher score/logit z^{T} are used as supervision for the student relevance classifier.

The teacher scores each query–chunk pair using the expanded query $\tilde{q}_i$, whereas the student receives only the base query q_i. This information asymmetry encourages the student to recover useful missing context through metadata selection and abstraction. The training objective combines hard-label binary cross-entropy with soft teacher distillation:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{\text{BCE}} + \alpha\mathcal{L}_{\text{KD}}, \tag{11}$$

where α balances hard-label learning and soft distillation. The binary cross-entropy loss is

$$\mathcal{L}_{\text{BCE}} = -y \log \sigma(z) - (1-y)\log(1-\sigma(z)), \tag{12}$$

where z is the student relevance logit and σ is the sigmoid function. The distillation term matches the teacher and student soft scores:

$$\mathcal{L}_{\text{KD}} = \text{KL}\left(\sigma(z^{\text{T}}/\tau) \parallel \sigma(z/\tau)\right), \tag{13}$$

where z^{T} is the teacher score/logit and τ is the temperature. The student encoder, topic centroids, and cached chunk topic distributions are kept fixed. We train only the metadata selector, abstraction module, and MLP relevance classifier.

3.4 Inference

At inference time, MCOMPASSRAG retrieves evidence without LLM calls. All chunk embeddings, topic distributions, and topic-enriched chunk representations are precomputed offline as indices for retrieval. For a given query, MCOMPASSRAG computes the query embedding, selects and abstracts relevant metadata from the bank, scores all cached chunks with the MLP classifier, and returns the top-k results. Algorithm 1 summarizes this procedure. Since topic extraction and chunk encoding are offline, online inference only requires lightweight metadata selection, abstraction, and scoring.

Algorithm 1 MCOMPASSRAG Inference

Require: Query q, precomputed chunk representations $\{\mathbf{r}_{c_j}\}_{j=1}^{N}$, metadata bank $\mathcal{M}$, topic centroids $\{\mathbf{t}_r\}_{r=1}^{K}$, selected metadata count L, top topics M, retrieved chunks k

Ensure: Retrieved chunk set $\mathcal{C}_k$

$\mathbf{e}_q \leftarrow f_\psi(q)$
// Metadata selection
for each metadata entry $\boldsymbol{\theta}_{c_i} \in \mathcal{M}$ **do**
  $\mathbf{m}_i \leftarrow \sum_{r=1}^{K} \theta_{c_i,r}\mathbf{t}_r$
  $a_i \leftarrow \mathbf{w}_s^\top[\mathbf{e}_q; \mathbf{m}_i] + b_s$
end for
$s_i \leftarrow \frac{\exp(a_i)}{\sum_{j=1}^{|\mathcal{M}|}\exp(a_j)}$
$\mathcal{S} \leftarrow \text{top-L}(\{s_i\})$
// Metadata abstraction
$\mathbf{H}^{(0)} \leftarrow [\boldsymbol{\theta}_{c_j}]_{j\in\mathcal{S}}$
$\hat{\boldsymbol{\theta}}_q \leftarrow \text{MeanPool}(\text{TransformerEnc}(\mathbf{H}^{(0)}))$
$\mathcal{T}_q \leftarrow \text{top-M}(\hat{\boldsymbol{\theta}}_q)$
$\mathbf{g}_q \leftarrow \sum_{r\in\mathcal{T}_q} \hat{\theta}_{q,r}\mathbf{t}_r$
$\mathbf{r}_q \leftarrow [\mathbf{e}_q; \mathbf{g}_q]$
// Retrieval
for each precomputed $\mathbf{r}_{c_j}$ **do**
  $z_j \leftarrow \text{MLP}_\phi([\mathbf{r}_q; \mathbf{r}_{c_j}])$
end for
$\mathcal{C}_k \leftarrow \text{top-k}_{c_j\in\mathcal{C}}(\{z_j\})$
return $\mathcal{C}_k$

4 Experiments and Results

4.1 Experimental Setup

Models and implementation. We use QWEN3-EMBEDDING-4B (Team, 2025) as the student encoder for query and chunk representations, and QWEN3-32B (Team, 2025) as both the LLM teacher for relevance supervision and the final answer generator. For baselines requiring LLM-based generation, planning, or selection, we use the same LLM scale for fair comparison. When a baseline requires reranking, we use QWEN3-RERANKER-4B (Zhang et al., 2025c). Closed-source API-based components are accessed through OpenRouter[2]. All experiments are run with access to 8 NVIDIA A100 80GB GPUs.

Topic metadata. We use CEMTM (Abaskohi et al., 2025) with QWEN3-EMBEDDING-4B as the topic modeling backbone. CEMTM is trained on WikiWeb2M (Burns et al., 2023) with 100 topics and is used across all experiments. In MCOMPASSRAG, we use only the CEMTM encoder to obtain chunk-level document-topic vectors and topic centroid embeddings. Since the LLM teacher also requires topic-aware representations, we additionally use a QWEN3-32B-based CEMTM variant for teacher-side topic modeling. We ablate the embedding backbone, topic model, and in-domain topic modeling in Appendices H, I, and J.

Benchmarks. We evaluate on seven benchmarks: SCI-DOCS (Cohan et al., 2020), LegalBench-RAG (Pipitone and Alami, 2024), Dragonball (Zhu et al., 2025), HotpotQA (Yang et al., 2018), SQuAD (Rajpurkar et al., 2016), DRBench (Abaskohi et al., 2026), and LongBenchV2 (Bai et al., 2025). For retrieval evaluation, we use SCI-DOCS, LegalBench-RAG, Dragonball, HotpotQA, SQuAD, and DRBench, which provide evidence annotations or links convertible to chunk-level labels. We use LongBenchV2 only for downstream evaluation, as it lacks chunk-level evidence labels. See ppendix B for more details.

Baselines. We compare against dense, struc-

[2] https://openrouter.ai/

Method	Dragonball			HotpotQA			SQuAD		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
RAPTOR	30.13±.41	39.40±.52	10.53±.29	45.43±.63	59.63±.58	13.70±.34	60.70±.71	32.77±.44	21.13±.39
Meta-Chunking-MSP	31.40±.38	40.20±.47	11.63±.31	55.70±.69	64.30±.62	17.97±.42	80.60±.58	41.97±.53	34.40±.49
Meta-Chunking-PPL	40.87±.45	42.80±.50	15.73±.36	66.77±.73	65.23±.64	21.40±.47	78.80±.62	41.37±.55	33.70±.51
DenseXRetrieval	2.27±.12	4.40±.18	0.09±.03	35.60±.56	43.17±.49	7.03±.21	61.53±.68	31.17±.46	19.83±.37
SAKI-RAG	32.90±.42	71.37±.66	25.40±.45	58.73±.70	55.60±.59	30.03±.52	87.17±.51	88.80±.43	78.93±.57
LLM	34.73±.39	76.53±.61	27.30±.43	62.63±.67	55.83±.55	33.50±.49	89.93±.46	91.63±.40	82.77±.52
LLM + 10 Topics	**40.83**±.34	**87.43**±.49	**34.17**±.38	**72.90**±.58	**59.33**±.51	**42.70**±.44	**94.10**±.33	**95.83**±.29	**89.50**±.36
MCompassRAG + 10 Topics	38.97±.36	82.80±.52	32.40±.40	70.17±.61	56.40±.48	40.63±.46	93.80±.35	95.37±.31	88.90±.38

Method	DRBench			LegalBench-RAG			SCI-DOCS		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
RAPTOR	24.13±.37	32.77±.44	8.20±.25	24.27±.35	32.23±.42	8.20±.24	88.63±.54	82.77±.50	80.37±.55
Meta-Chunking-MSP	30.60±.42	36.13±.47	12.30±.31	28.30±.39	36.10±.45	11.07±.29	90.47±.49	83.53±.48	82.10±.52
Meta-Chunking-PPL	36.30±.49	37.57±.51	16.17±.34	32.70±.43	37.53±.48	13.57±.32	21.07±.36	17.60±.31	3.57±.15
DenseXRetrieval	18.40±.31	25.37±.38	5.43±.19	19.53±.33	24.93±.36	5.13±.18	86.00±.57	79.33±.53	74.67±.60
SAKI-RAG	37.47±.46	62.30±.61	28.23±.43	31.23±.41	46.30±.52	19.27±.36	86.53±.50	92.27±.43	84.30±.51
LLM	41.53±.44	68.43±.57	32.27±.41	33.93±.39	50.40±.49	22.13±.35	89.37±.45	95.10±.34	87.47±.43
LLM + 10 Topics	**50.27**±.39	**83.17**±.46	**43.30**±.37	**40.10**±.34	**59.47**±.43	**29.70**±.31	**94.67**±.30	**99.50**±.12	**92.50**±.28
MCompassRAG + 10 Topics	47.97±.41	78.57±.49	41.20±.39	38.40±.36	55.10±.45	27.90±.33	94.13±.32	99.03±.15	92.10±.29

Table 1: Retrieval performance across six benchmarks, averaged over three runs. ± values denote standard deviation. **Bold** = best; underline = second-best; shaded rows indicate MCompassRAG. LLM-based rows are inference-time oracle upper bounds. Detailed k=1, 3, 5 results are in Appendix D.

tured, long-context, and LLM-based RAG baselines: DenseXRetrieval (Chen et al., 2024), Meta-Chunking with PPL and MSP variants (Zhao et al., 2025b), RAPTOR (Sarthi et al., 2024), ReflectiveRAG (Verma et al., 2026), DF-RAG (Khan et al., 2026), SAKI-RAG (Tao et al., 2025), REFRAG (Lin et al., 2025), PageIndex (Zhang et al., 2025a), A-RAG (Du et al., 2026), Context-1 (Bashir et al., 2026), and a long-context Qwen3-32B baseline. For retrieval evaluation, we include DenseXRetrieval, Meta-Chunking, RAPTOR, SAKI-RAG, and LLM retrievers, evaluating both topic-free and topic-guided LLM variants. The remaining baselines are evaluated only on downstream performance because they primarily target generation, decoding, reranking, or context-use efficiency rather than standalone retrieval. Refer to Appendix B for more details.

Training and evaluation. We train MCompassRAG separately for each benchmark, using synthetic training data when retrieval labels are unavailable or insufficient; for DRBench and LongBenchV2, we train on EDR-200 (Prabhakar et al., 2025) and LongBenchV1 (Bai et al., 2024), respectively. We train only the metadata selector, abstraction module, and MLP classifier, while keeping all encoders and cached topic representations fixed. Retrieval quality is measured by Recall, Precision, and Information Efficiency (IE), with IE@k = Precision@k × Recall@k, averaged over $k \in \{1, 3, 5\}$ and three runs. Downstream performance is evaluated with task-appropriate metrics: Accuracy, F1, ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang* et al., 2020). For fair comparison across chunk granularities, retrieved chunks are added in ranked order until a fixed token budget is reached (1K). Full training hyperparameters, inference, and evaluation settings are provided in Appendix C.

4.2 Comparison with Retrieval Baselines

Table 1 reports retrieval performance across all six benchmarks. MCompassRAG with 10 topic signals **consistently outperforms all baselines** across every benchmark and metric. The gains are most pronounced on harder, multi-hop benchmarks: on DRBench, MCompassRAG achieves an IE of 47.97 versus 37.47 for the strongest non-LLM baseline (SAKI-RAG), and on LegalBench-RAG it similarly leads on all three metrics. On SCI-DOCS and SQuAD, where retrieval is comparatively easier, MCompassRAG still matches or exceeds all baselines with comfortable margins. Notably, **MCompassRAG closely approaches the LLM + 10 Topics oracle**, which invokes a full LLM at retrieval time, **while requiring no**

	Final Performance							Average Cost	
	HotpotQA	LongBench v2		DRBench	Dragonball				
Method	F1 ↑	F1 ↑	Acc ↑	F1 ↑	R-L ↑	MTR ↑	BRT ↑	Tok/Q ↓	Lat. (ms) ↓
Efficient RAG Methods									
Dense X Retrieval	60.9	26.4	28.6	46.8	0.248	0.269	0.548	2759	112
Meta-Chunking-PPL	64.5	29.7	31.8	50.7	0.272	0.292	0.571	2394	95
RAPTOR	63.1	28.3	30.4	49.1	0.264	0.285	0.563	3183	145
ReflectiveRAG	67.4	31.5	33.4	53.4	0.303	0.325	0.604	3527	161
DF-RAG	66.2	30.2	32.3	52.1	0.291	0.313	0.592	4843	484
SAKI-RAG	68.6	32.6	34.5	55.2	0.314	0.336	0.619	5584	925
REFRAG	73.6	37.5	39.4	60.4	0.354	0.371	0.650	7800	720
Long-Context Methods									
PageIndex	78.7	41.9	43.6	65.8	0.372	0.394	0.682	53 883	4408
A-RAG	74.9	38.7	40.4	62.4	0.347	0.369	0.655	14 625	2557
Context-1	76.1	40.1	41.8	64.1	0.359	0.382	0.669	20 430	3026
LLM	72.9	36.9	38.8	59.3	0.352	0.362	0.642	41 058	3388
Ours									
MCompassRAG	71.8	35.8	35.7	58.9	0.333	0.355	0.635	4126	174

Table 2: Downstream generation performance and efficiency across four benchmarks. R-L = ROUGE-L, MTR = METEOR, BRT = BERTScore. Tok/Q = tokens per query; Lat. = end-to-end latency.

inference-time LLM calls: the IE gap is under 1 point on SCI-DOCS (94.13 vs. 94.67) and SQuAD (93.80 vs. 94.10), and within 2–3 points on the remaining benchmarks. The consistent gap between the topic-free LLM and LLM + 10 Topics rows further confirms that **topic metadata carries substantial guidance value beyond raw chunk embeddings**, which MCompassRAG exploits efficiently through lightweight distillation rather than runtime LLM inference. Appendix K provides qualitative examples illustrating how topic signals resolve retrieval failures that dense similarity cannot handle.

4.3 Downstream Performance and Efficiency

Table 2 compares downstream generation quality and efficiency across all methods. Among **efficient RAG methods**, MCompassRAG achieves competitive generation quality while remaining one of the most efficient systems. With only **4,126 tokens per query and 174 ms end-to-end latency**, MCompassRAG is substantially cheaper than SAKI-RAG (5,584 tok, 925 ms) and REFRAG (7,800 tok, 720 ms), the two strongest efficient baselines in generation quality. The performance gap between MCompassRAG and these methods is largely attributable to their use of LLM-based reranking or context selection at inference time, which filters out noisy evidence before generation at the cost of additional latency. MCompassRAG recovers much of this quality through topic-guided retrieval alone, without any post-retrieval LLM filtering. Although MCompassRAG requires training, this is a one-time cost rather than an inference-time overhead; moreover, Table 4.4 shows that the trained retriever can **generalize across datasets** when trained on a general dataset like MS Marco (Nguyen et al., 2016), further amortizing this cost even when switching to new corpora. **Compared to long-context methods**, MCompassRAG operates at over 10× fewer tokens than PageIndex and the LLM baseline, while delivering generation scores within a reasonable margin. The remaining gap reflects the fact that long-context methods can exploit all available evidence in the document, whereas MCompassRAG is constrained to a fixed retrieval budget; the key finding is that topic-guided coarse retrieval recovers most of the evidence quality of expensive long-context methods at a fraction of the cost.

4.4 Ablation Studies

The Effect of Abstraction and Selection Policy. Table 3 (blue rows) shows that removing either the abstraction module or the selection policy consistently lowers IE, with the largest drop when both are removed. The selection policy identifies query-relevant metadata, while the abstraction module denoises and compresses the selected topic distributions into a usable query-side signal. Without selection, abstraction receives weaker metadata; without abstraction, selected topics remain a noisy raw mixture. Their complementary roles explain

Method	Dragonball			DRBench			LegalBench-RAG			SCI-DOCS		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
MCompassRAG	38.97	82.80	32.40	47.97	78.57	41.20	38.40	55.10	27.90	94.13	99.03	92.10
W/O Abst.	38.03	82.27	31.90	47.50	77.93	40.23	37.93	54.70	27.47	93.27	98.63	91.87
W/O Select Pol.	38.53	80.30	31.37	48.20	74.93	38.70	38.20	53.27	26.53	93.87	97.13	91.30
W/O Abst. + W/O Select Pol.	37.47	80.83	31.13	45.93	75.63	38.27	37.30	53.90	26.80	92.40	97.87	91.00
MSMarco (Nguyen et al., 2016)	36.20	78.37	29.30	44.53	73.03	35.73	36.03	52.10	24.60	91.20	96.37	88.97
CLaRa (He et al., 2025)	35.30	77.27	28.10	43.47	71.23	33.27	35.23	51.03	23.30	90.27	95.40	86.90

Table 3: Ablation study and training data generalizability across four benchmarks. The blue shows the full MCompassRAG model and its component ablations. Pink rows show MCompassRAG trained on out-of-domain datasets (MSMarco and CLaRa) rather than the target benchmark, evaluating generalizability of the distillation pipeline without in-domain training data. See Appendix E for the results over all six benchmarks.

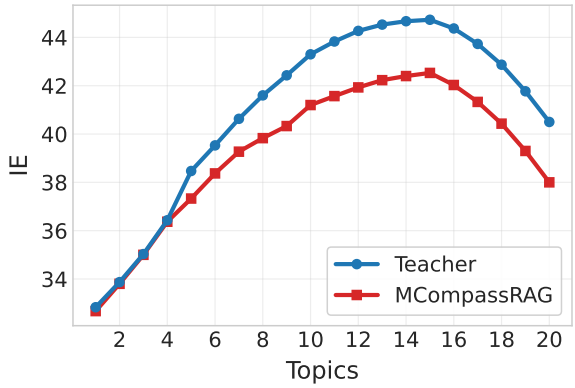


Figure 3: IE as a function of the number of selected topics on DRBench for the full MCompassRAG pipeline, comparing the LLM teacher and student retriever.

why the full MCompassRAG pipeline performs best across benchmarks.

Effect of the Number of Metadata Topic. Figure 3 analyzes how the number of selected topics affects IE on DRBench for the full MCompassRAG pipeline. Performance peaks at an intermediate topic count, around 15 topics, and degrades when too many topics are provided. This suggests that topic metadata is useful as a compact semantic guide, but excessive topic information can dilute the original query–chunk signal. The teacher consistently outperforms the student, as it receives richer per-topic representations, while the student relies on an abstracted topic summary. However, the gap remains modest around the optimal topic range, indicating that the selection and abstraction modules preserve most of the useful teacher signal for the lightweight retriever. We provide the full analysis across Dragonball and DRBench, including all ablation variants, in Appendix F.

Training Data Generalizability. The pink rows in Table 3 show MCompassRAG trained on MSMarco (Nguyen et al., 2016) and CLaRa (He et al., 2025) without any access to target-benchmark data. Despite having no in-domain supervision, **both variants substantially outperform all non-LLM baselines** from Table 1 across every benchmark. The performance gap relative to in-domain training is modest in most settings, indicating that the distillation pipeline learns transferable retrieval behavior rather than overfitting to benchmark-specific patterns. This is practically important: MCompassRAG **does not require labeled in-domain data** to deliver strong topic-guided retrieval, making it straightforward to deploy in new domains without additional annotation.

5 Conclusion and Future Works

We introduced MCompassRAG, a metadata-guided retrieval framework that enriches coarse chunk representations with topic-level signals and trains a lightweight student retriever through LLM-teacher distillation, enabling topic-aware retrieval without inference-time LLM calls. Across six retrieval benchmarks, MCompassRAG improves information efficiency by 8.24% on average over the strongest non-LLM baseline while running at over 5× lower latency compared to strong LLM-based baselines. Ablation studies confirm that both the metadata selection policy and the abstraction module are necessary, and that the distillation pipeline generalizes well without in-domain training data. Several promising directions build on this work: jointly optimizing the topic model and retriever end-to-end could better align topic representations and further close the student–teacher gap; developing approximate selection strategies would improve scalability to very large corpora; and integrating MCompassRAG into iterative deep research agents is a natural next step, where efficiency gains compound across multiple retrieval rounds.

Limitations

MCOMPASSRAG has a few limitations worth noting. First, the quality of topic-guided retrieval is directly dependent on the quality of the underlying topic model: poorly trained or misaligned topic representations will produce uninformative metadata signals. This creates a dependency on having a reliable topic model for the target corpus, which may require additional effort in low-resource or highly specialized domains. Second, MCOMPASSRAG introduces several hyperparameters — the number of topics for the topic modeling model K, the number of selected metadata entries from the memory bank L, the number of metadata to use for retrieval M, and the number of retrieved chunks k — whose interactions are non-trivial to tune. As shown in Section 4.4, performance is sensitive to the number of topics passed to the model, and finding the right operating point requires careful validation. Third, the current topic enrichment strategy represents each chunk and query as a weighted sum of topic centroid embeddings, which is a lossy compression: combining multiple topic vectors into a single aggregated vector discards the individual structure of each topic signal. As the topic count increases, this aggregation becomes increasingly noisy. Developing richer representations that preserve per-topic structure without inflating the vector dimensionality — for instance through sparse or cross-attention-based topic integration — is an important direction for future work.

Ethical Considerations

Potential Risks. MCOMPASSRAG is designed to improve the efficiency and quality of evidence retrieval in RAG systems, but like any retrieval framework it inherits the risks associated with the underlying corpus. If the document collection contains biased, harmful, or factually incorrect content, MCOMPASSRAG may retrieve and surface such content more effectively, potentially amplifying its downstream impact. Topic models trained on biased corpora may also learn skewed topic representations that systematically favor or suppress certain types of evidence. Users deploying MCOMPASSRAG in high-stakes domains such as legal or medical question answering should carefully audit both the corpus and the topic model for such biases before deployment.

Use of AI Assistance. During the preparation of this manuscript, we used large language models for paraphrasing and improving the writing quality of the paper, and for generating icon images used in figures. All ideas, technical contributions, experimental design, and conclusions are entirely original and the work of the authors.

Reproducibility Statement. We describe all architectural components, training objectives, hyperparameters, and dataset splits in Section 4.1. Training data is synthesized from publicly available benchmarks using prompts provided in Appendix A. Upon publication, we will release our code and checkpoints under a permissive open-source license. All external datasets and models are used according to their original licenses and terms of use. All benchmarks, datasets, and baseline artifacts used in this paper are publicly available, and our use of them is limited to research evaluation, consistent with their intended use and original access conditions. Our released code and checkpoints are intended for research purposes and for reproducing the experiments in this paper.

Use of Existing Artifacts and Data. We use only publicly available benchmarks, datasets, models, and baseline artifacts, and our use is limited to research evaluation. We do not collect new user data or attempt to identify individuals. Some source corpora may contain naturally occurring named entities or web-derived text from the original datasets; we follow the datasets' original licenses, preprocessing, and access conditions. All benchmarks, datasets, and baseline methods used in this paper are publicly available, and our released code and checkpoints are intended for research and reproducibility purposes.

References

Amirhossein Abaskohi, Tianyi Chen, Miguel Muñoz-Mármol, Curtis Fox, Amrutha Varshini Ramesh, Étienne Marcotte, Xing Han Lù, Nicolas Chapados, Spandana Gella, Christopher Pal, Alexandre Drouin, and Issam H. Laradji. 2026. DRBench: A realistic benchmark for enterprise deep research. In *The Fourteenth International Conference on Learning Representations*.

Amirhossein Abaskohi, Raymond Li, Chuyuan Li, Shafiq Joty, and Giuseppe Carenini. 2025. CEMTM: Contextual embedding-based multimodal topic modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*,

pages 11675–11692, Suzhou, China. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hammad Bashir, Kelly Hong, Patrick Jiang, and Zhiyi Shi. 2026. Chroma context-1: Training a self-editing search agent. Technical report, Chroma.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. In *The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. 2026. A-rag: Scaling agentic retrieval-augmented generation via hierarchical retrieval interfaces. *Preprint*, arXiv:2602.03442.

Zheng Fang, Yulan He, and Rob Procter. 2024. CWTM: Leveraging contextualized word embeddings from BERT for neural topic modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4273–4286, Torino, Italia. ELRA and ICCL.

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, Toronto, Canada. Association for Computational Linguistics.

Jie He, Richard He Bai, Sinead Williamson, Jeff Z. Pan, Navdeep Jaitly, and Yizhe Zhang. 2025. Clara: Bridging retrieval and generation with continuous latent reasoning. *Preprint*, arXiv:2511.18659.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

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 (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Saadat Hasan Khan, Spencer Hong, Jingyu Wu, Kevin Lybarger, Youbing Yin, Erin Babinsky, and Daben Liu. 2026. DF-RAG: Query-aware diversity for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2873–2894, Rabat, Morocco. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Raymond Li, Amirhossein Abaskohi, Chuyuan Li, Gabriel Murray, and Giuseppe Carenini. 2026. Improving neural topic modeling with semantically-grounded soft label distributions. *Preprint*, arXiv:2602.17907.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiaoqiang Lin, Aritra Ghosh, Bryan Kian Hsiang Low, Anshumali Shrivastava, and Vijai Mohan. 2025. Refrag: Rethinking rag based decoding. *Preprint*, arXiv:2509.01092.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Maxime Louis, Thibault Formal, Hervé Déjean, and Stéphane Clinchant. 2026. OSCAR: Online soft compression for RAG. In *The Fourteenth International Conference on Learning Representations*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

OpenAI. 2024. Introducing GPT-4o and more tools to ChatGPT free users. `https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/`.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Nicholas Pipitone and Ghita Houir Alami. 2024. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain. *Preprint*, arXiv:2408.10343.

Akshara Prabhakar, Roshan Ram, Zixiang Chen, Silvio Savarese, Frank Wang, Caiming Xiong, Huan Wang, and Weiran Yao. 2025. Enterprise deep research: Steerable multi-agent deep research for enterprise analytics. *Preprint*, arXiv:2510.17797.

Suchith Chidananda Prabhu, Bhavyajeet Singh, Anshul Mittal, Siddarth Asokan, Shikhar Mohan, Deepak Saini, Yashoteja Prabhu, Lakshya Kumar, Jian Jiao, Amit S, Niket Tandon, Manish Gupta, Sumeet Agarwal, and Manik Varma. 2025. MOGIC: Metadata-infused oracle guidance for improved extreme classification. In *Forty-second International Conference on Machine Learning*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Wenyu Tao, Xiaofen Xing, Zeliang Li, and Xiangmin Xu. 2025. SAKI-RAG: Mitigating context fragmentation in long-document RAG via sentence-level attention knowledge integration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1195–1213, Suzhou, China. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Akshay Verma, Swapnil Gupta, Siddharth Pillai, Prateek Sircar, and Deepak Gupta. 2026. ReflectiveRAG: Rethinking adaptivity in retrieval-augmented generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 377–384, Rabat, Morocco. Association for Computational Linguistics.

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, Singapore. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Mingtian Zhang, Yu Tang, and PageIndex Team. 2025a. Pageindex: Next-generation vectorless, reasoning-based rag. *PageIndex Blog*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. 2025b. Deep research: A survey of autonomous research agents. *Preprint*, arXiv:2508.12752.

Xuechen Zhang, Koustava Goswami, Samet Oymak, Jiasi Chen, and Nedim Lipka. 2026. Smartchunk retrieval: Query-aware chunk compression with planning for efficient document RAG. In *The Fourteenth International Conference on Learning Representations*.

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. 2025c. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Jihao Zhao, Zhiyuan Ji, Zhaoxin Fan, Hanyu Wang, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025a. MoC: Mixtures of text chunking learners for retrieval-augmented generation system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5172–5189, Vienna, Austria. Association for Computational Linguistics.

Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025b. Meta-chunking: Learning text segmentation and semantic completion via logical perception. *Preprint*, arXiv:2410.12788.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4).

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *The Twelfth International Conference on Learning Representations*.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. DeepResearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431, Suzhou, China. Association for Computational Linguistics.

Weichao Zhou, Jiaxin Zhang, Hilaf Hasson, Anu Singh, and Wenchao Li. 2024. HyQE: Ranking contexts with hypothetical query embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13014–13032, Miami, Florida, USA. Association for Computational Linguistics.

Kunlun Zhu, Yifan Luo, Dingling Xu, Yukun Yan, Zhenghao Liu, Shi Yu, Ruobing Wang, Shuo Wang, Yishan Li, Nan Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. RAGEval: Scenario specific RAG evaluation dataset generation framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8520–8544, Vienna, Austria. Association for Computational Linguistics.

A Prompts Used for Training

This appendix lists the prompts used during training. Prompt A.1 is used to generate base and expanded queries from training chunks. The next prompt, Prompt A.2, is used by the LLM teacher to assign relevance labels to query–chunk pairs during distillation.

Prompt A.1: Query Expansion

You are given three consecutive chunks from a document: the previous chunk, the target chunk, and the next chunk.
Your task has two steps.
Step 1: Generate a base query. Write a natural user question that requires information from the target chunk to answer. The question should not directly copy the answer from the chunk, and it should not reveal the answer.
Step 2: Generate an expanded query. Rewrite the base query by adding useful background context from the previous and next chunks. The expanded query should make the information need clearer, but it must not reveal the answer or include direct answer hints. Use only background context that helps specify the topic, setting, entities, or surrounding discussion.
Input:
Previous chunk: {previous_chunk}
Target chunk: {target_chunk}
Next chunk: {next_chunk}
Output format:
Base query: {base query}
Expanded query: {expanded query}

Prompt A.2: Teacher Relevance Labeling

You are given a question and a candidate knowledge chunk. Decide whether the chunk contains information that is useful for answering the question.
Mark the chunk as relevant only if it provides direct or supporting evidence needed to answer the question. Do not mark a chunk as relevant based only on vague topical similarity.
Question: {expanded_query}
Candidate chunk: {candidate_chunk}
Output only one number:
1 = relevant
0 = not relevant

B Benchmark and Baseline Details

B.1 Benchmark Dataset Details

We evaluate MCOMPASSRAG on seven benchmarks spanning scientific, legal, open-domain multi-hop, reading comprehension, enterprise deep research, and long-context tasks. Table 4 summarizes key statistics.

SCI-DOCS (Cohan et al., 2020). SCI-DOCS is a comprehensive evaluation suite for scientific document embeddings, covering seven document-level tasks ranging from citation prediction and document classification to recommendation, and including tens of thousands of examples of anonymized user signals of document relatedness. It was introduced alongside the SPECTER model to address the limitation that prior evaluations of scientific document representations focused on small datasets over a limited set of tasks, where extremely high AUC scores were already achievable. The corpus consists of scientific paper abstracts, which are naturally multi-topic and stylistically homogeneous, making it a natural testbed for topic-guided retrieval.

LegalBench-RAG (Pipitone and Alami, 2024). LegalBench-RAG is the first benchmark designed specifically to evaluate the retrieval step of RAG pipelines in the legal domain. It is constructed by retracing the context used in LegalBench queries back to their original locations within the legal corpus, resulting in 6,858 query-answer pairs over a corpus of over 79 million characters, entirely human-annotated by legal experts. The dataset covers a diverse range of legal documents including NDAs, M&A agreements, commercial contracts, and privacy policies. The benchmark demands precise, minimal snippet retrieval rather than broad document recall, making it an especially challenging test of fine-grained retrieval.

Dragonball (Zhu et al., 2025). Dragonball (Diverse RAG Omni-Benchmark for All domains) is released as part of the RAGEval framework. It contains 6,711 questions meticulously designed to reflect the complexity and specificity of their domains, covering finance, legal, and medical scenarios in both Chinese and English. The framework introduces three novel keypoint-based metrics—Completeness, Hallucination, and Irrelevance—to evaluate generated responses by distilling standard answers into 3–5 key points encompassing indispensable factual information and final conclusions. Dragonball's multilingual and multi-domain construction stresses retrieval systems operating over heterogeneous, topically distinct evidence pools.

HotpotQA (Yang et al., 2018). HotpotQA contains 113k Wikipedia-based question-answer pairs featuring four key properties: questions require finding and reasoning over multiple supporting documents; questions are diverse and unconstrained by any knowledge base schema; sentence-level sup-

Dataset	Domain	Language	#Queries (eval)	Corpus Size (#docs)	Avg. Doc. Len. (tokens)	Multi-hop
SCI-DOCS	Scientific	EN	1,000	25k	7,955	✗
LegalBench-RAG	Legal	EN	6,858	714	27.13k	✗
Dragonball	Finance/Legal/Medical	EN+ZH	6,711	2,311	11,436	✗
HotpotQA	Open-domain	EN	113k	105k	1,247	✓
SQuAD	Open-domain	EN	107,785	536	2,303	✗
DRBench	Enterprise	EN	1,093	1,093	1,089	✓
LongBenchV2	Multi-task	EN	503	503	59.38k	✓

Table 4: Statistics of the seven benchmark datasets used in our evaluation. "Avg. Doc. Len." reports average document length in characters. "#Queries (eval)" refers to the number of queries used in our experiments. "Multi-hop" indicates whether the benchmark requires cross-document reasoning.

porting facts are provided for reasoning supervision; and a category of factoid comparison questions tests the ability to extract and compare relevant facts across entities. Sentence-level supporting fact annotations make HotpotQA directly usable for chunk-level retrieval evaluation; its multi-hop structure requires retrievers to surface evidence distributed across distinct document segments.

SQuAD (Rajpurkar et al., 2016). SQuAD contains 107,785 question-answer pairs on 536 Wikipedia articles, where the answer to every question is a text span from the corresponding reading passage. It covers a wide range of topics from musical celebrities to abstract concepts. Unlike HotpotQA, SQuAD questions are largely single-passage answerable, providing a complementary single-hop retrieval axis in our evaluation.

DRBench (Abaskohi et al., 2026). DRBench evaluates AI agents on complex, open-ended deep research tasks in enterprise settings, requiring agents to identify supporting facts from both the public web and private company knowledge bases. Each task is grounded in realistic user personas and enterprise context, spanning a heterogeneous search space that includes productivity software, cloud file systems, emails, chat conversations, and the open web. The benchmark targets report generation in enterprise deep research settings, comprising 100 tasks with a total of 1,093 sub-questions.

LongBenchV2 (Bai et al., 2025). LongBenchV2 consists of 503 challenging multiple-choice questions, with contexts ranging from 8k to 2M words, across six major task categories: single-document QA, multi-document QA, long in-context learning, long-dialogue history understanding, code repository understanding, and long structured data understanding. Data was collected from nearly 100 highly educated individuals with diverse professional backgrounds. LongBenchV2 is used exclusively for downstream generation evaluation, as it does not provide chunk-level evidence labels that can serve as retrieval ground truth.

B.2 Baseline Method Details

We compare against eleven baselines. We describe each method's core methodology below, along with which part of the pipeline—retrieval or generation—it primarily targets.

DenseXRetrieval (Chen et al., 2024). DenseXRetrieval introduces the *proposition* as a novel retrieval unit for dense retrieval. Propositions are defined as atomic expressions within text, each encapsulating a distinct factoid and presented in a concise, self-contained natural language format. A fine-tuned generation model called the Propositionizer—trained via a two-step distillation process—decomposes passages into their constituent propositions at indexing time.

Meta-Chunking (PPL and MSP) (Zhao et al., 2025b). Meta-Chunking leverages LLMs' logical perception capabilities to identify optimal text segment boundaries, moving beyond fixed-size and similarity-based chunking. It defines a meta-chunk granularity between sentences and paragraphs, consisting of sentences with deep linguistic logical connections. Two adaptive uncertainty-driven strategies are proposed: *Perplexity (PPL) Chunking*, which identifies boundaries by analyzing the context perplexity distribution of an LLM—splitting at points of certainty and keeping intact at uncertainty; and *Margin Sampling (MSP) Chunking*, which uses LLMs to perform binary classification on whether consecutive sentences should be segmented based on the probability difference from margin sampling. Additionally, a global information compensation mechanism—comprising a two-stage hierarchical summary generation process and a three-stage chunk rewriting procedure—preserves se-

mantic integrity and contextual coherence across chunks.

RAPTOR (Sarthi et al., 2024). RAPTOR introduces the novel approach of recursively embedding, clustering, and summarizing chunks of text to construct a tree with differing levels of summarization from the bottom up. At inference time, retrieval operates across all tree levels, enabling queries to be answered by combining evidence from fine-grained passages and their higher-level summaries.

SAKI-RAG (Tao et al., 2025). SAKI-RAG addresses context fragmentation in long-document RAG via two core components: (1) the SentenceAttnLinker, which constructs a semantically enriched knowledge repository by modeling inter-sentence attention relationships; and (2) the Dual-Axis Retriever, which expands and filters candidate chunks along both the semantic similarity and contextual relevance dimensions.

ReflectiveRAG (Verma et al., 2026). ReflectiveRAG addresses two persistent inefficiencies in standard RAG: static top-k retrieval regardless of evidence sufficiency, and context redundancy from semantically overlapping retrieved passages. Current methods—fixed top-k retrieval, cross-encoder reranking, or policy-based iteration—rely on static heuristics or costly reinforcement learning, failing to assess evidence sufficiency or reduce redundancy. ReflectiveRAG introduces a Self-Reflective Retrieval (SRR) module that uses a compact language model to iteratively evaluate whether retrieved evidence is sufficient or requires further query reformulation, alongside a Noise Removal (NR) module that scores and filters retrieved chunks by relevance minus redundancy.

DF-RAG (Khan et al., 2026). DF-RAG systematically incorporates diversity into the retrieval step to improve performance on complex, reasoning-intensive QA benchmarks. It builds upon the Maximal Marginal Relevance framework to select information chunks that are both relevant to the query and maximally dissimilar from each other. A key innovation is its ability to optimize the level of diversity for each query dynamically at test time without requiring any additional fine-tuning or prior information.

REFRAG (Lin et al., 2025). REFRAG targets generation-side efficiency by exploiting block-diagonal attention patterns that arise from low inter-passage semantic similarity among retrieved chunks. It uses a compress–sense–expand framework: a lightweight encoder compresses each retrieved chunk into compact embeddings fed directly to the decoder; an RL-trained policy selectively determines which chunks require full token-level expansion; and the decoder operates over a substantially shorter effective input.

PageIndex (Zhang et al., 2025a). PageIndex replaces the standard chunk–embed–vector search pipeline with a hierarchical tree index built from documents, using an LLM to reason over that tree—analogous to how a human expert scans a table of contents. Rather than passive similarity lookup, PageIndex performs active tree search, with the LLM navigating document structure across multiple reasoning steps. Retrieval happens inline during the model's reasoning process, allowing the system to begin streaming immediately without a blocking retrieval gate before the first token.

A-RAG (Du et al., 2026). A-RAG proposes an agentic RAG framework that exposes hierarchical retrieval interfaces directly to the language model. Unlike existing methods that either retrieve passages in a single shot and concatenate them into input, or predefine a workflow and prompt the model to execute it step-by-step, A-RAG allows the model to adapt the retrieval strategy based on the specific task, choose different interaction strategies, and decide when sufficient evidence has been gathered to provide an answer. A-RAG satisfies three principles of agentic autonomy—Autonomous Strategy, Iterative Execution, and Interleaved Tool Use—making it a truly agentic framework.

Context-1 (Bashir et al., 2026). Chroma Context-1 is a 20B parameter agentic search model derived from GPT-OSS-20B (OpenAI, 2025) that achieves retrieval performance comparable to frontier-scale LLMs at a fraction of the cost and up to 10$\times$ faster inference speed. It is designed to be used as a subagent in conjunction with a frontier reasoning model: given a query, it produces a ranked list of documents relevant to satisfying the query. The model is trained to decompose queries into subqueries, iteratively search a corpus, and selectively edit its own context to free capacity for further exploration. A key mechanism is self-editing context management, in which the agent actively discards retrieved passages deemed irrele-

vant as the context window fills, preventing context rot during long-horizon multi-hop retrieval.

Long-context LLM Baseline (Qwen3-32B). As a strong non-retrieval baseline, we use Qwen3-32B in a full long-context setting where the model both identifies relevant evidence and generates the final answer from the complete document context. This baseline directly tests whether a state-of-the-art long-context LLM can substitute retrieval entirely. We additionally evaluate two LLM retriever variants: a *topic-free* variant, where the LLM selects relevant chunks without topic guidance, and a *topic-guided* variant, where the LLM is provided with topic metadata as part of its retrieval prompt.

C Training and Implementation Details

Models and implementation. We use QWEN3-EMBEDDING-4B (Team, 2025) as the student encoder for query and chunk representations. We ablate the choice of embedding backbone in Appendix H. For LLM-based teacher supervision, we use QWEN3-32B (Team, 2025) to generate relevance judgments and soft teacher scores. The same QWEN3-32B model is also used as the final generator for answer generation. For fair comparison, we use the same LLM scale for all baselines that require LLM-based generation, planning, or selection. When a baseline requires a reranker, we use QWEN3-RERANKER-4B (Zhang et al., 2025c). For closed-source API-based components, we access models through OpenRouter[3]. All experiments are run with access to 8 NVIDIA A100 80GB GPUs.

Topic modeling. For topic metadata, we use CEMTM (Abaskohi et al., 2025) with QWEN3-EMBEDDING-4B as the embedding backbone. We train CEMTM on WikiWeb2M (Burns et al., 2023) using the same hyperparameters as the original CEMTM setup and use the resulting topic model across our experiments. During the MCOMPASS-RAG pipeline, we only use the encoder component of CEMTM to obtain chunk-level document-topic vectors and topic centroid embeddings. Since the LLM teacher also requires topic-aware representations, we additionally use a QWEN3-32B-based CEMTM variant for teacher-side topic modeling. We set the number of topics to $K = 100$. We ablate the choice of topic model in Appendix I. We additionally evaluate an in-domain variant in which the topic model is trained directly on the target corpus; results are provided in Appendix J.

Training details. For each benchmark, MCOMPASSRAG is trained separately using its corresponding training split. When a benchmark does not provide a sufficiently large training set, we use 10% of the available data for synthetic training data construction. For DRBench (Abaskohi et al., 2026) and LongBenchV2 (Bai et al., 2025), which are smaller and do not provide suitable retrieval training labels, we train using EDR-200 (Prabhakar et al., 2025) and LongBenchV1 (Bai et al., 2024), respectively. For each dataset, we sample 2,000 training chunks and generate 10 synthetic queries per chunk, resulting in 20,000 query–chunk pairs before negative sampling. We train the metadata selector, abstraction module, and MLP relevance classifier while keeping the student encoder, topic centroids, and cached chunk-topic distributions fixed. Unless otherwise specified, all hyperparameters follow our default setting: AdamW (Loshchilov and Hutter, 2019) with learning rate 2×10^{-5}, batch size 16, weight decay 0.01, dropout 0.1, and 3 training epochs. The distillation temperature is set to $\tau = 1.0$, and the loss interpolation coefficient is set to $\alpha = 0.5$. For generation, we use temperature $\tau = 0.7$ and top-$p = 0.9$; for teacher relevance scoring, we use temperature $\tau = 0.0$ to obtain deterministic judgments.

Evaluation. Because the compared methods use different chunk granularities, evaluating all systems with a fixed number of retrieved chunks can be unfair: the same top-k may correspond to very different amounts of retrieved text. We therefore use two complementary evaluation protocols. For retrieval quality, we report Recall, Precision, and Information Efficiency (IE), where $\text{IE@k} = \text{Precision@k} \times \text{Recall@k}$. These metrics are computed at $k \in \{1, 3, 5\}$ and averaged over three runs. For downstream evaluation, we use task-appropriate generation metrics, including Accuracy, F1, ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang* et al., 2020), depending on the benchmark. To ensure fairness in downstream comparisons, retrieved chunks are added in ranked order until a fixed token budget is reached (1K), so each method provides the generator with the same maximum amount of evidence regardless of its chunk size. This protocol evaluates retrieval methods under comparable evidence budgets while still allowing each method to use its

[3] https://openrouter.ai/

own native chunking strategy. We use $L = 50$ and $M = 10$ in our experiments.

D Retrieval Performance at Different Cutoffs

Tables 5, 6, and 7 report retrieval performance at cutoffs k=1, k=3, and k=5, respectively, across all six benchmarks. As expected, both precision and recall increase monotonically with k for all methods, since retrieving more documents provides greater coverage of relevant passages. The relative ordering of methods remains consistent across all cutoffs: MCOMPASSRAG outperforms all non-oracle baselines at every depth while staying within a narrow margin of the LLM + 10 Topics oracle, which relies on an LLM at retrieval time. This consistency demonstrates that the gains from topic-guided retrieval are not specific to any particular cutoff, but reflect a robust improvement in retrieval quality across the full range of evaluation settings reported here.

E Ablation Results on All Benchmarks

Table 8 presents the complete ablation study and training data generalizability results across all six benchmarks, extending the four-benchmark summary reported in the Table 3.

F Detailed Analysis on the Effect Metadata Topics

Figure 4 extends the topic-count analysis to Dragonball and DRBench across four ablation variants. The same trend observed in the main paper holds consistently: IE improves as the number of selected topics increases up to an intermediate range, typically around 10–15 topics, and then decreases as additional topics introduce noise. This supports our design choice of using a compact set of query-relevant topic signals rather than exposing the retriever to a large number of corpus-level topic vectors. The teacher–student gap also varies with the number of topics. At low topic counts, both models receive limited topic information, leading to a small gap. At intermediate counts, topic metadata is most informative, and the teacher benefits more from its richer per-topic representation. At high counts, both models suffer from topic overload, causing their performance to converge again. Across ablations, the full MCOMPASSRAG pipeline achieves the strongest student IE and the smallest teacher–student gap, showing that both metadata selection and abstraction are important for making topic information usable by the lightweight student retriever.

G Effect of Topic Granularity of Topic Model

Table 9 reports retrieval performance as a function of the number of topics K in the underlying topic model. Two consistent patterns emerge across all three benchmarks. First, **performance peaks at $K = 100$ and degrades monotonically as K increases beyond this point**. At very high granularities ($K = 500$–2000), topic representations become increasingly fine-grained and sparse, making each topic centroid less representative of a coherent semantic direction. As a result, the weighted aggregation of topic centroids produces chunk and query representations that are noisier and harder to match reliably. Second, **the student–teacher gap is largest at $K = 100$ and nearly vanishes at high K**. At $K = 100$, the LLM teacher can exploit the richer and more semantically coherent per-topic structure to outperform the student, which receives only a compressed topic summary. At $K \geq 500$, both the teacher and the student suffer equally from the degraded topic quality, and their performance converges. Together, these results suggest that **a moderate topic granularity of $K = 100$ strikes the best balance** between topic coherence and coverage, and we use this setting across all experiments in the main paper. This finding is complementary to the analysis in Section 4.4, which studied the effect of how many topic signals are passed to the model at inference time: here we show that the quality of those signals, determined by K, is equally important. Even with an optimal number of passed topics, overly fine-grained or coarse topic models will degrade retrieval quality.

H Sensitivity to the Embedding Model

To assess whether MCOMPASSRAG depends on a specific embedding backbone, we evaluate its retrieval performance with different embedding models while keeping the rest of the pipeline unchanged. Table 10 reports results on three representative benchmarks: Dragonball, LegalBench-RAG, and SCI-DOCS. We compare the main QWEN3-EMBEDDING-4B configuration against smaller and larger Qwen embedding models, a projected QWEN3-EMBEDDING-0.6B variant, BAAI/BGE-M3, and ALL-MINILM-L6-V2.

Method	Dragonball			HotpotQA			SQuAD		
	IE@1 ↑	*Prec.*@1 ↑	*Rec.*@1 ↑	*IE*@1 ↑	*Prec.*@1 ↑	*Rec.*@1 ↑	*IE*@1 ↑	*Prec.*@1 ↑	*Rec.*@1 ↑
RAPTOR	2.74	36.40	7.53	5.40	55.63	9.70	5.40	29.77	18.13
Meta-Chunking-MSP	3.21	37.20	8.63	8.42	60.30	13.97	12.24	38.97	31.40
Meta-Chunking-PPL	5.07	39.80	12.73	10.65	**61.23**	17.40	11.78	38.37	30.70
DenseXRetrieval	0.00	1.08	0.32	1.19	39.17	3.03	4.74	28.17	16.83
SAKI-RAG	15.31	68.37	22.40	13.43	51.60	26.03	65.15	85.80	75.93
LLM	17.87	73.53	24.30	15.29	51.83	29.50	70.70	88.63	79.77
LLM + 10 Topics	**26.32**	**84.43**	**31.17**	**21.41**	55.33	**38.70**	**80.30**	**92.83**	**86.50**
MCompassRAG + 10 Topics	23.46	79.80	29.40	19.19	52.40	36.63	79.35	92.37	85.90

Method	DRBench			LegalBench-RAG			SCI-DOCS		
	IE@1 ↑	*Prec.*@1 ↑	*Rec.*@1 ↑	*IE*@1 ↑	*Prec.*@1 ↑	*Rec.*@1 ↑	*IE*@1 ↑	*Prec.*@1 ↑	*Rec.*@1 ↑
RAPTOR	1.38	29.27	4.70	1.52	29.23	5.20	62.51	80.27	77.87
Meta-Chunking-MSP	2.87	32.63	8.80	2.67	33.10	8.07	64.50	81.03	79.60
Meta-Chunking-PPL	4.32	34.07	12.67	3.65	34.53	10.57	0.16	15.10	1.07
DenseXRetrieval	0.42	21.87	1.93	0.47	21.93	2.13	55.45	76.83	72.17
SAKI-RAG	14.54	58.80	24.73	7.04	43.30	16.27	73.43	89.77	81.80
LLM	18.68	64.93	28.77	9.07	47.40	19.13	78.68	92.60	84.97
LLM + 10 Topics	**31.71**	**79.67**	**39.80**	**15.08**	**56.47**	**26.70**	**89.19**	**99.10**	**90.00**
MCompassRAG + 10 Topics	28.30	75.07	37.70	12.97	52.10	24.90	88.03	98.25	89.60

Table 5: Retrieval performance at depth k=1 across six benchmarks (IE @1 ↑, Precision @1 ↑, Recall @1 ↑). **Bold** = best; underline = second-best. MCompassRAG rows are shaded. LLM and LLM + 10 Topics are oracle upper bounds that use an LLM at retrieval time.

Overall, stronger embedding models improve retrieval quality, with Qwen3-Embedding-8B yielding the best results and Qwen3-Embedding-4B remaining very close while being more efficient. The projected Qwen3-Embedding-0.6B variant consistently improves over the unprojected version, suggesting that alignment with the main embedding space is important for using topic metadata effectively. Although all-MiniLM-L6-v2 is substantially smaller than the Qwen-based encoders, it still achieves competitive performance compared with several baselines reported in Table 1. This indicates that the gains of MCompassRAG do not come only from using a strong embedding backbone; the metadata selection and abstraction mechanism itself provides useful retrieval guidance across different encoder choices.

I Sensitivity to the Topic Model

To evaluate whether MCompassRAG depends on a particular topic model, we replace the topic encoder while keeping the rest of the retrieval pipeline fixed. Table 11 compares four topic modeling approaches: ETM (Dieng et al., 2020), CWTM (Fang et al., 2024), SoftLTM (Li et al., 2026), and CEMTM. ETM represents topics and words in a shared embedding space and learns topics through the interaction between topic embeddings and word embeddings, making it a natural baseline for embedding-space topic guidance. CWTM extends this direction by incorporating contextualized representations when estimating document-topic distributions, which generally provides more semantically informed topic assignments than purely lexical topic models. SoftLTM refers to the soft-label topic modeling approach of Li et al. (2026), which trains topic models using semantically grounded soft label distributions produced by a language model. Since SoftLTM does not directly provide topic centroids in the same form required by MCompassRAG, we construct each topic centroid by averaging the embedding vectors of its top topic words.

CEMTM is the main topic model used in our experiments. Compared with the alternatives, CEMTM provides topic representations that are directly aligned with the retriever embedding space and produces document-topic vectors using stronger semantic supervision. As shown in Table 11, CEMTM achieves the best overall retrieval performance. However, CWTM and SoftLTM remain competitive, with CWTM slightly outperforming SoftLTM across the three datasets. This

Method	Dragonball			HotpotQA			SQuAD		
	IE@3 ↑	*Prec.*@3 ↑	*Rec.*@3 ↑	*IE*@3 ↑	*Prec.*@3 ↑	*Rec.*@3 ↑	*IE*@3 ↑	*Prec.*@3 ↑	*Rec.*@3 ↑
RAPTOR	3.78	38.65	9.78	7.45	58.63	12.70	6.53	32.02	20.38
Meta-Chunking-MSP	4.29	39.45	10.88	10.74	63.30	16.97	13.87	41.22	33.65
Meta-Chunking-PPL	6.30	42.05	14.98	13.10	**64.23**	20.40	13.38	40.62	32.95
DenseXRetrieval	0.03	2.65	1.07	2.54	42.17	6.03	5.80	30.42	19.08
SAKI-RAG	17.41	70.62	24.65	15.85	54.60	29.03	68.84	88.05	78.18
LLM	20.12	75.78	26.55	17.82	54.83	32.50	74.54	90.88	82.02
LLM + 10 Topics	**28.97**	**86.68**	**33.42**	**24.32**	58.33	**41.70**	**84.38**	**95.08**	**88.75**
MCompassRAG + 10 Topics	25.97	82.05	31.65	21.96	55.40	39.63	83.41	94.62	88.15

Method	DRBench			LegalBench-RAG			SCI-DOCS		
	IE@3 ↑	*Prec.*@3 ↑	*Rec.*@3 ↑	*IE*@3 ↑	*Prec.*@3 ↑	*Rec.*@3 ↑	*IE*@3 ↑	*Prec.*@3 ↑	*Rec.*@3 ↑
RAPTOR	2.34	31.90	7.32	2.35	31.48	7.45	65.51	82.14	79.75
Meta-Chunking-MSP	4.03	35.26	11.43	3.65	35.35	10.32	67.55	82.91	81.47
Meta-Chunking-PPL	5.62	36.70	15.30	4.72	36.78	12.82	0.50	16.98	2.94
DenseXRetrieval	1.11	24.50	4.55	1.06	24.18	4.38	58.28	78.70	74.05
SAKI-RAG	16.80	61.42	27.36	8.44	45.55	18.52	76.68	91.64	83.67
LLM	21.21	67.56	31.40	10.62	49.65	21.38	82.04	94.47	86.84
LLM + 10 Topics	**34.91**	**82.30**	**42.42**	**17.00**	**58.72**	**28.95**	**91.33**	**99.40**	**91.88**
MCompassRAG + 10 Topics	31.33	77.69	40.33	14.76	54.35	27.15	90.41	98.84	91.47

Table 6: Retrieval performance at depth k=3 across six benchmarks (IE @3 ↑, Precision @3 ↑, Recall @3 ↑). **Bold** = best; underline = second-best. MCompassRAG rows are shaded.

suggests that MCompassRAG is not tied to a single topic model; rather, its main requirement is that the topic model provides meaningful document-topic distributions and topic centroids that can be mapped into the retriever embedding space.

J Topic Model Domain Adaptation: Training on Target Corpus

In the main experiments, we use a topic model trained on WikiWeb2M to provide a general-purpose set of topic centroids and document-topic vectors. While this setting tests whether MCompassRAG can rely on a broadly trained topic model, some benchmarks contain domain-specific terminology and evidence structures that may not be fully captured by a general corpus. We therefore evaluate an in-domain variant in which the topic model is trained directly on the target corpus of each benchmark, while keeping the rest of the MCompassRAG pipeline unchanged.

Table 12 compares the default WikiWeb2M-trained topic model with target-corpus topic models on Dragonball, LegalBench-RAG, and SCI-DOCS. Training the topic model on the target corpus improves performance across all three datasets, with larger gains on LegalBench-RAG and Dragonball, where domain-specific terminology, entities, and narrative structure are especially important. However, the gains are moderate rather than dramatic, showing that MCompassRAG does not require retraining the topic model for every new corpus. This is important for practical deployment: a general-purpose topic model can already provide useful metadata guidance, while in-domain topic modeling can be used as an optional enhancement when sufficient target-corpus data and training budget are available.

K Qualitative Analysis

We present two qualitative examples to illustrate how MCompassRAG resolves retrieval failures that dense similarity cannot handle: a definitional ambiguity case from LegalBench-RAG and an embedding-space analysis from Dragonball Finance.

LegalBench-RAG: definitional ambiguity in M&A agreements. Figure 5 illustrates a concrete retrieval example from LegalBench-RAG that exposes the core failure mode of dense retrieval and how MCompassRAG resolves it. The query asks for the definition of *Superior Proposal* in an M&A acquisition agreement whose §6.03 region contains several topically adjacent clauses: a no-shop obligation (C1), the definition of *Acquisition Proposal*

Method	Dragonball			HotpotQA			SQuAD		
	IE@5 ↑	*Prec.*@5 ↑	*Rec.*@5 ↑	*IE*@5 ↑	*Prec.*@5 ↑	*Rec.*@5 ↑	*IE*@5 ↑	*Prec.*@5 ↑	*Rec.*@5 ↑
RAPTOR	6.16	43.15	14.28	12.09	64.63	18.70	9.09	36.52	24.88
Meta-Chunking-MSP	6.76	43.95	15.38	15.92	69.30	22.97	17.44	45.72	38.15
Meta-Chunking-PPL	9.07	46.55	19.48	18.54	**70.23**	26.40	16.90	45.12	37.45
DenseXRetrieval	0.02	8.15	0.20	5.79	48.17	12.03	8.23	34.92	23.58
SAKI-RAG	21.90	75.12	29.15	21.23	60.60	35.03	76.52	92.55	82.68
LLM	24.93	80.28	31.05	23.42	60.83	38.50	82.52	95.38	86.52
LLM + 10 Topics	**34.58**	**91.18**	**37.92**	**30.69**	64.33	**47.70**	**92.86**	**99.58**	**93.25**
MCompassRAG + 10 Topics	31.29	86.55	36.15	28.02	61.40	45.63	91.83	99.12	92.65

Method	DRBench			LegalBench-RAG			SCI-DOCS		
	IE@5 ↑	*Prec.*@5 ↑	*Rec.*@5 ↑	*IE*@5 ↑	*Prec.*@5 ↑	*Rec.*@5 ↑	*IE*@5 ↑	*Prec.*@5 ↑	*Rec.*@5 ↑
RAPTOR	4.67	37.15	12.57	4.30	35.98	11.95	71.72	85.89	83.50
Meta-Chunking-MSP	6.76	40.51	16.68	5.91	39.85	14.82	73.85	86.66	85.22
Meta-Chunking-PPL	8.62	41.95	20.55	7.15	41.28	17.32	1.39	20.73	6.70
DenseXRetrieval	2.92	29.75	9.80	2.55	28.68	8.88	64.15	82.45	77.80
SAKI-RAG	21.74	66.67	32.61	11.52	50.05	23.02	83.39	95.39	87.42
LLM	26.68	72.81	36.65	14.01	54.15	25.88	88.98	98.22	90.59
LLM + 10 Topics	**41.74**	**87.55**	**47.67**	**21.15**	**63.22**	**33.45**	**95.62**	**100.00**	**95.62**
MCompassRAG + 10 Topics	37.80	82.94	45.58	18.63	58.85	31.65	95.22	100.00	95.22

Table 7: Retrieval performance at depth k=5 across six benchmarks (IE @5 ↑, Precision @5 ↑, Recall @5 ↑). **Bold** = best; underline = second-best. MCompassRAG rows are shaded.

(C2), a board recommendation withdrawal clause (C4), and a termination fee clause (C5). Dense retrieval assigns the highest cosine similarity to C2 (0.81) rather than the gold chunk C3 (0.78), ranking the wrong definition first. The failure arises because C2 and C3 share substantial surface vocabulary ("bona fide," "majority," "Acquisition," "outstanding Shares"), causing their embeddings to occupy nearby positions in the retriever space. Cosine similarity cannot identify which latent topic of a chunk matches the query, nor distinguish a clause that *defines* what counts as an acquisition proposal from one that *evaluates* whether a proposal is superior.

MCompassRAG recovers the gold chunk by activating two topic signals identified by the metadata selector as compatible with the query embedding: T-A, capturing the fiduciary-out and board determination frame ("more favorable," "financial advisor," "board determines in good faith"), and T-B, capturing the majority threshold frame ("majority of outstanding Shares," "bona fide written Acquisition Proposal"). The selector simultaneously suppresses signals associated with solicitation restrictions (T-C) and merger consideration (T-D), which are prominent in the neighboring chunks but orthogonal to the query's information need. The abstraction module pools T-A and T-B into a compact query-side topic vector aligned with C3's own topic representation, and the MLP classifier assigns C3 a relevance score of 0.89 versus 0.57 for C2, promoting the correct definition to rank 1 without any inference-time LLM call. This disambiguation was learned through the teacher–student asymmetry in Section 3.3: the LLM teacher, given the expanded query framing the information need in terms of board determination and financial-advisor consultation, labels C3 as relevant and C2 as not, training the student to recover the same judgment through topic metadata alone.

Dragonball Finance: topic-guided separation in embedding space. Figure 6 visualizes the effect of topic enrichment on a Dragonball Finance example in which the query asks for a summary of Sparkling Clean Housekeeping Services' sustainability and social responsibility efforts in 2019. The eight retrieval candidates span the full thematic range of the company's corporate governance report: board composition (C1), executive remuneration (C2), risk management (C3), financial highlights (C4), shareholder structure (C5), internal audit (C6), and two surface-overlap distractors whose vocabulary partially overlaps with the gold chunk: a compliance and anti-corruption clause

Method	Dragonball			HotpotQA			SQuAD		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
MCompassRAG	38.97	82.80	32.40	70.17	56.40	40.63	93.80	95.37	88.90
W/O Abst.	38.03	82.27	31.90	69.30	56.20	40.20	93.03	94.93	88.37
W/O Select Pol.	38.53	80.30	31.37	70.07	55.93	39.07	93.53	93.80	87.93
W/O Abst. + W/O Select Pol.	37.47	80.83	31.13	68.27	55.97	39.43	92.50	94.10	87.47
MSMarco (Nguyen et al., 2016)	36.20	78.37	29.30	66.23	55.57	36.40	91.40	93.13	85.43
CLaRa (He et al., 2025)	35.30	77.27	28.10	64.67	55.30	34.53	90.60	92.20	83.63

Method	DRBench			LegalBench-RAG			SCI-DOCS		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
MCompassRAG	47.97	78.57	41.20	38.40	55.10	27.90	94.13	99.03	92.10
W/O Abst.	47.50	77.93	40.23	37.93	54.70	27.47	93.27	98.63	91.87
W/O Selection Pol.	48.20	74.93	38.70	38.20	53.27	26.53	93.87	97.13	91.30
W/O Abst. + W/O Selection Pol.	45.93	75.63	38.27	37.30	53.90	26.80	92.40	97.87	91.00
MSMarco (Nguyen et al., 2016)	44.53	73.03	35.73	36.03	52.10	24.60	91.20	96.37	88.97
CLaRa (He et al., 2025)	43.47	71.23	33.27	35.23	51.03	23.30	90.27	95.40	86.90

Table 8: Ablation study and training data generalizability across six benchmarks. The top block (blue) shows the full MCompassRAG model and its component ablations. Pink rows show MCompassRAG trained on out-of-domain datasets (MSMarco and CLaRa) rather than the target benchmark, evaluating generalizability of the distillation pipeline without in-domain training data.

(C7, which shares the phrase "corporate citizenship") and a strategic outlook statement (C8, which shares "long-term value creation").

In the raw embedding space (Figure 6a), the query and the gold CSR chunk are already relatively proximate, yet several hard negatives remain in the same neighbourhood, reflecting the broad semantic overlap that coarse governance-report language introduces. After topic enrichment (Figure 6b), the query–gold alignment tightens substantially: the metadata selector activates the CSR topic centroid for the query and the gold chunk's own topic distribution loads on the same signal, pulling the two representations into close alignment while the hard negatives, whose dominant topic vectors correspond to governance, finance, and risk, drift away. The surface-overlap distractors C7 and C8 are particularly informative: despite sharing specific phrases with the gold chunk, their topic distributions do not load on the CSR centroid and therefore receive lower relevance scores from the MLP classifier, confirming that MCompassRAG's disambiguation operates at the level of latent topic structure rather than lexical overlap.

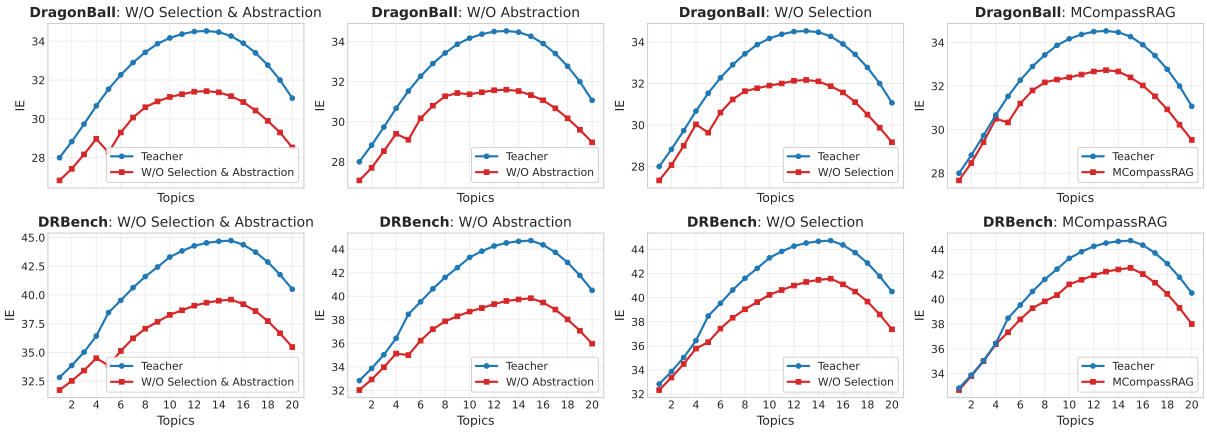


Figure 4: IE as a function of the number of topics passed to the model, comparing the teacher and student (MCompassRAG) across four ablation variants on Dragonball and DRBench. Each column removes one component of the metadata selection and abstraction pipeline.

		SCI-DOCS			LegalBench-RAG			Dragonball		
K	**Method**	Recall ↑	Precision ↑	IE ↑	Recall ↑	Precision ↑	IE ↑	Recall ↑	Precision ↑	IE ↑
50	MCompassRAG	88.83	93.37	86.87	36.23	51.40	26.30	36.73	78.53	30.47
	LLM	92.43	98.13	91.90	37.70	55.43	27.83	38.43	82.60	32.07
100	MCompassRAG	94.13	99.03	92.10	38.40	55.10	27.90	38.97	82.80	32.40
	LLM	98.30	99.63	98.03	40.10	59.47	29.70	40.83	87.43	34.17
500	MCompassRAG	86.53	89.63	83.47	35.30	49.87	25.27	35.60	74.90	28.77
	LLM	87.20	90.40	84.03	35.57	50.30	25.43	35.97	75.37	28.97
1000	MCompassRAG	84.80	87.10	81.17	34.60	48.47	24.57	34.63	72.83	27.80
	LLM	85.13	87.47	81.50	34.73	48.67	24.67	34.83	73.07	27.90
2000	MCompassRAG	83.40	85.23	79.60	34.03	47.43	24.10	33.90	71.27	27.07
	LLM	83.57	85.40	79.83	34.10	47.53	24.17	34.00	71.40	27.13

Table 9: Effect of topic model granularity (K) on retrieval performance across three datasets. Results are reported for MCompassRAG and LLM-based methods.

Embedding Backbone	**Dragonball**			**LegalBench-RAG**			**SCI-DOCS**		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
QWEN3-EMBEDDING-0.6B	34.83	74.16	28.74	34.27	48.86	23.93	90.68	95.41	88.17
QWEN3-EMBEDDING-0.6B + PROJECTION	36.38	77.34	30.21	36.06	51.73	25.68	92.04	96.77	89.86
BAAI/BGE-M3	35.91	76.08	29.76	35.14	50.31	24.83	92.57	97.18	90.82
ALL-MINILM-L6-V2	29.64	64.23	23.47	28.92	41.79	18.94	84.23	89.27	79.36
QWEN3-EMBEDDING-4B	38.97	82.80	32.40	38.40	55.10	27.90	94.13	99.03	92.10
QWEN3-EMBEDDING-8B	**39.43**	**83.46**	**32.91**	**38.88**	**55.77**	**28.36**	**94.39**	**99.18**	**92.47**

Table 10: Embedding-backbone ablation for MCOMPASSRAG on three representative retrieval benchmarks. Results report IE↑, Precision↑, and Recall↑, averaged over retrieval cutoffs k=1, 3, 5. The QWEN3-EMBEDDING-4B row corresponds to the main configuration used in Table 1; other rows show expected trends before running the full ablation. **Bold** = best; underline = second-best.

Topic Model	Dragonball			LegalBench-RAG			SCI-DOCS		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
ETM	33.74	71.28	27.31	32.86	47.14	22.76	89.42	94.36	86.91
SoftLTM	36.83	78.64	30.57	36.38	52.19	25.91	92.71	97.46	90.49
CWTM	37.28	79.31	30.94	36.76	52.63	26.24	93.08	97.91	90.96
CEMTM	**38.97**	**82.80**	**32.40**	**38.40**	**55.10**	**27.90**	**94.13**	**99.03**	**92.10**

Table 11: Topic-model ablation for MCompassRAG on three representative retrieval benchmarks. Results report IE↑, Precision↑, and Recall↑, averaged over retrieval cutoffs k=1, 3, 5. CEMTM is the main configuration used in Table 1.

Topic Model Training Corpus	Dragonball			LegalBench-RAG			SCI-DOCS		
	IE↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑	*IE*↑	*Prec.*↑	*Rec.*↑
WikiWeb2M	38.97	82.80	32.40	38.40	55.10	27.90	94.13	99.03	92.10
Target Corpus	**39.26**	**83.71**	**32.83**	**40.18**	**57.36**	**29.64**	**94.82**	**99.21**	**92.86**

Table 12: Effect of training the topic model on the target corpus. Results report IE↑, Precision↑, and Recall↑, averaged over retrieval cutoffs k=1, 3, 5. The WikiWeb2M row corresponds to the main MCompassRAG configuration, while the Target Corpus row trains the topic model on the corresponding benchmark corpus before running the same retrieval pipeline.

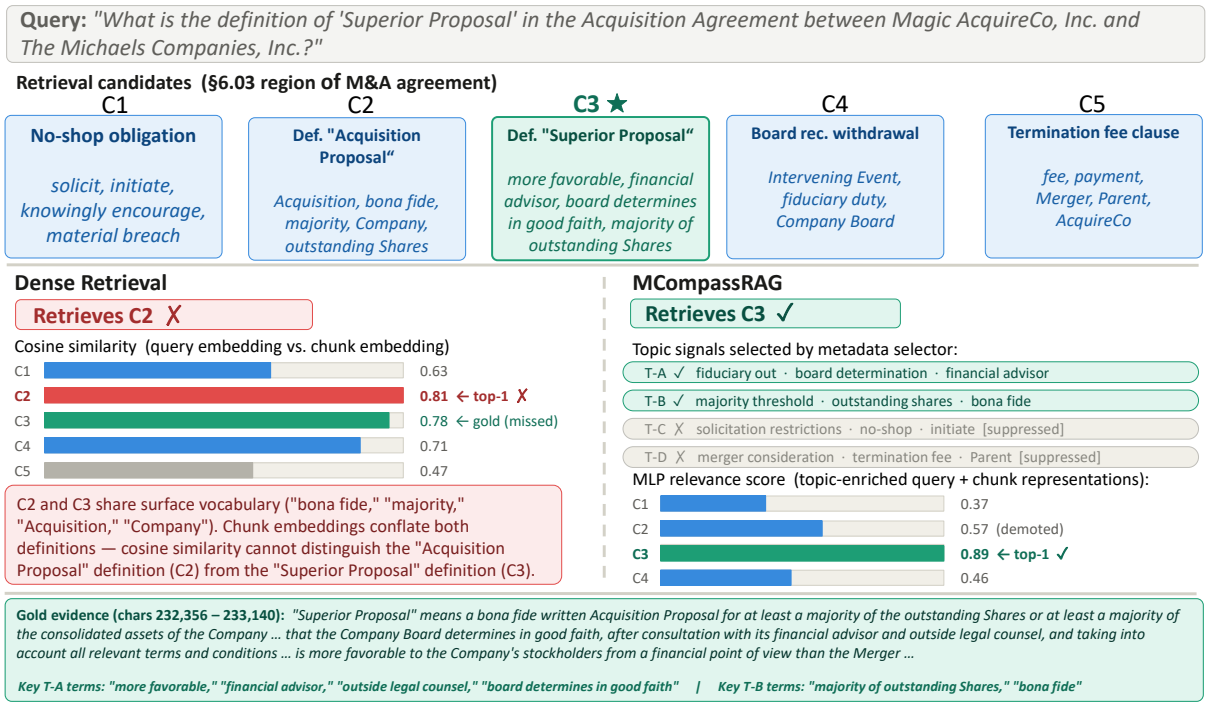


Figure 5: Qualitative retrieval comparison on LegalBench-RAG for a query about the definition of *Superior Proposal* in an M&A acquisition agreement. **Top**: five retrieval candidates from the §6.03 region; the gold chunk (C3, teal border) competes against four topically adjacent clauses sharing substantial surface vocabulary. **Bottom left**: dense retrieval ranks C2 (*Acquisition Proposal* definition) above C3 due to overlapping tokens, missing the gold chunk at rank 1. **Bottom right**: MCompassRAG activates topic signals T-A (fiduciary out / board determination) and T-B (majority threshold), suppresses T-C and T-D, and promotes C3 to rank 1 via the MLP scorer (0.89 vs. 0.57 for C2).

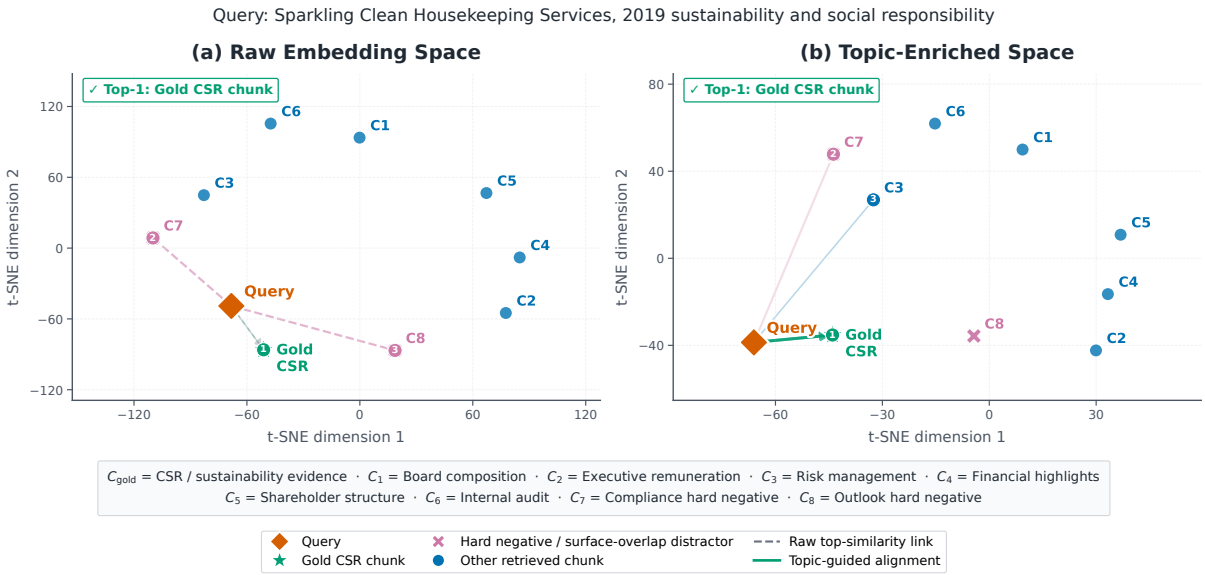


Figure 6: t-SNE visualization of chunk embeddings for a Dragonball Finance query on Sparkling Clean Housekeeping Services' 2019 sustainability efforts. Chunks cover eight aspects of the corporate governance report: board composition (C1), executive remuneration (C2), risk management (C3), financial highlights (C4), shareholde structure (C5), internal audit (C6), compliance and anti-corruption (C7), and strategic outlook (C8); C7 and C8 are surface-overlap distractors that share phrases with the gold CSR chunk. **(a) Raw embedding space**: the query and gold chunk are proximate but several hard negatives occupy the same neighbourhood. **(b) Topic-enriched space**: topic enrichment tightens the query–gold alignment while pushing all hard negatives, including the surface-overlap distractors C7 and C8, away from the query.