TopicSummRAG: A Topic-Enhanced RAG model for Improving Query-Focused Summarization from Long Documents

Anonymous authors
Paper under double-blind review

Abstract

Despite larger context windows, large language models (LLMs) continue to struggle with answering queries over long, unstructured documents. Retrieval-Augmented Generation (RAG) mitigates this limitation by retrieving relevant text before generation; however, its effectiveness depends critically on how documents are segmented into retrievable units. Existing chunking strategies—such as fixed-size or sliding windows—often ignore topical coherence, frequently merging unrelated content or fragmenting coherent discourse, which degrades retrieval precision and downstream generation quality. We propose TopicSummRAG, a framework that aligns retrieval units with the latent topical structure of documents. TopicSummRAG first segments documents into contextually coherent topical chunks using a boundary-supervision-free contrastive text segmentation model, and then summarises each segment to form compact retrieval metadata. The segmentation component is evaluated independently on the QMSum and TIAGE benchmarks, where it consistently improves boundary detection and placement over strong baselines. During retrieval, segment-level summaries are matched against the query, and an entropy- and dominance-based filtering strategy adaptively selects relevant segments by measuring the concentration of relevance mass, avoiding brittle fixed cutoffs. We evaluate TopicSummRAG on ODSum-Story, ODSum-Meeting, and QMSum across multiple retrievers (BM25, SBERT, Situated) and LLMs (Qwen3-8B, LLaMA-3.1-8B, Gemma-3-12B). TopicSummRAG yields improvements of up to 13% ROUGE-1, 25% ROUGE-2, and 20% ROUGE-L, alongside 10–15 point gains in LENS and up to 10-point improvements in BLANC. These results demonstrate that topic-aware segmentation and adaptive retrieval substantially improve factual grounding, coherence, and retrieval robustness, providing a retriever- and model-agnostic framework for long-document query-focused summarisation and question answering with RAG.

1 Introduction

Despite advances in large language models (LLMs) and their ability to process extended contexts, query-focused information access over long, multi-topic documents remains a challenging task. Tasks such as query-focused summarisation (QFS) or question answering often involve long reports or conversational transcripts that span multiple topics and discourse phases. While LLMs can ingest long contexts, they lack an inherent mechanism to isolate query-relevant regions, frequently resulting in diffuse attention, reduced factual precision, or hallucinated content. Retrieval-Augmented Generation (RAG) mitigates this issue by retrieving relevant text before generation, but its effectiveness critically depends on document segmentation. Most RAG systems rely on fixed-size chunks, sliding windows, or sentence-level units that ignore latent topical structure. Such heuristics merge unrelated content or fragment coherent discourse, degrading retrieval quality—an issue amplified in long documents with abrupt shifts and gradual topical drift.

The notion of a topic varies substantially across NLP subfields. Latent topic models define topics as distributions of words, while dense retrieval encodes topics implicitly in embedding spaces. We instead define topics as contiguous discourse units—spans of text that are internally coherent and separated by semantic shifts. In this work, we propose TopicSummRAG, a framework that aligns retrieval units with the latent topical organisation of long documents. TopicSummRAG segments documents into contextually coherent topical

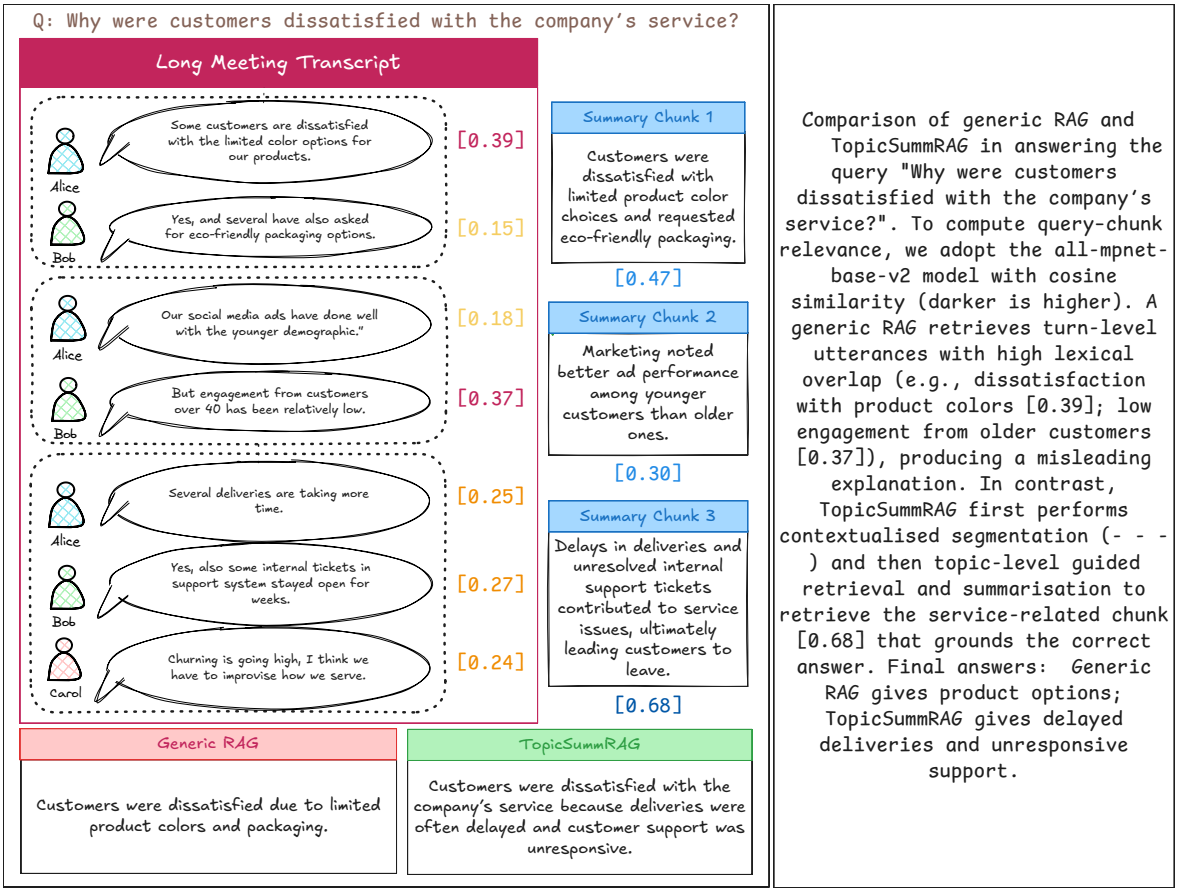

Figure 1: Comparison of generic RAG and TopicSummRAG

spans using a learned segmentation model and summarises each segment to form compact retrieval metadata. During inference, queries are matched against these segment-level summaries to select discourse-consistent regions for generation.

Figure 1 presents a motivating example illustrating the limitations of generic RAG and the advantages of our approach. The task is to answer queries over a long meeting transcript containing a multi-party conversation, where each speaker's contribution is treated as a turn. For a query such as "Why were customers dissatisfied with the company's service?", a generic RAG retriever ranks individual turns based on semantic similarity, often favouring utterances with high lexical overlap—even when they are only tangentially related to the underlying service issues. Treating each turn as an independent retrieval unit thus surfaces contextually misaligned evidence, which can mislead the LLM during generation. In contrast, TopicSummRAG retrieves coherent discourse segments corresponding to service dissatisfaction (e.g., delayed deliveries and unresponsive support), while excluding lexically similar but topically irrelevant content.

We evaluate our segmentation model on TIAGE (Xie et al., 2021) and QMSum (Zhong et al., 2021), and evaluate TopicSummRAG on multiple QFS benchmarks, including ODSum-Story, ODSum-Meeting (Rao et al., 2024), and QMSum, against strong RAG baselines. Across datasets, TopicSummRAG consistently improves retrieval accuracy and factual faithfulness while substantially reducing hallucinations. Overall, we contribute (i) a novel topic segmentation model for long documents and (ii) TopicSummRAG, a retrieval-augmented summarisation framework that improves RAG performance through topic-aware chunking and segment-level summaries, significantly outperforming state-of-the-art methods on query-focused long-document summarisation benchmarks.

2 Related Work

Query-Focused Summarization QFS aims to generate summaries that directly satisfy an explicit information need rather than providing a generic overview. Early extractive and abstractive approaches relied on lexical salience and query–document similarity (Baumel et al., 2016; 2018), which perform well on short or topically concentrated texts but degrade on long, multi-topic documents. Domain-specific QFS systems in areas such as biomedicine and law emphasise grounded evidence and factual consistency (Sankhavara & Majumder, 2020; Conrad et al., 2009). Recent query-based RAG methods (e.g., QB-RAG) improve alignment by matching queries against structured query representations (Yang et al., 2025). However, most approaches remain constrained by their retrieval units, typically operating at sentence or turn granularity or over truncated contexts, which disconnects retrieved evidence from the broader discourse structure required for coherent summarisation.

Retrieval-Augmented Generation and Retrieval Units Retrieval-augmented generation (RAG) grounds generation in external evidence to reduce hallucinations (Lewis et al., 2020; Brown et al., 2025). Extensions incorporating structured or graph-based retrieval improve global reasoning but still rely on heuristic segmentation and weak topic alignment (Edge et al., 2025; Hong et al., 2025). Consequently, RAG-based QFS systems face three recurring issues: (i) a precision–coverage trade-off as retrieval units grow; (ii) semantic misalignment between lexically similar but topically distinct spans; and (iii) scalability constraints when costly segmentation is applied at inference. Recent work shows that segmentation quality critically affects retrieval and downstream performance (Wang et al., 2025). TopicSummRAG addresses these limitations by using discourse-coherent segments and segment-level summaries as retrieval metadata. While segmentation is query-agnostic, retrieval is explicitly query-conditioned through summary-based matching and selection.

Topic Segmentation and Chunking Fixed-size and sentence-level chunking are efficient but often disrupt discourse coherence (Yepes et al., 2024). Classical unsupervised methods (e.g., TextTiling, C99) rely on lexical overlap and struggle with contextual coherence (Hearst, 1997). Recent approaches leverage contextual embeddings and neural models to better capture discourse structure (Song et al., 2022; Koshorek et al., 2018; Nair et al., 2023). Very recent work optimises chunking for RAG using learned or LLM-informed partitioning (Duarte et al., 2024; Zhao et al., 2025; Günther et al., 2025), but typically treats segmentation and retrieval as loosely coupled.

TopicSummRAG jointly addresses segmentation, retrieval, and summarisation by learning discourse-aware topical segments, summarising them into compact retrieval units, and applying adaptive, dominance-based selection. This design explicitly controls retrieval granularity, enabling coherent and recall-oriented query-focused retrieval over long documents.

3 Methodology

We study the problem of query-focused retrieval and summarisation over long, multi-topic documents. Given a document $D = \{u_1, u_2, \ldots, u_n\}$ composed of n textual units—such as sentences, paragraphs, or dialogue turns—and a user query q, the goal is to retrieve a compact set of evidence chunks $C = \{c_1, c_2, \ldots, c_m\}$ that collectively satisfy the information need expressed in q, and to generate a grounded, coherent response. A key challenge in this setting is that long documents often exhibit both abrupt topic transitions and gradual topical drift, while lacking explicit markers of topical boundaries. As a result, fixed-size or heuristic chunking strategies often fail to align retrieval units with the underlying semantic structure of the document, resulting in imprecise retrieval and degraded generation quality.

To address these challenges, we propose TopicSummRAG (Figure 6), a two-stage framework that explicitly aligns retrieval units with the latent topical organisation of long documents. The framework first segments a document into coherent topical spans and then performs topic-guided retrieval and summarisation over these spans. The first stage segments a long document into coherent topical spans, i.e., contiguous sequences of units that share a dominant topic (Section 3.1). At this stage, the aim is to detect the smallest textual block that is topically coherent and cohesive, which may span across sentences, paragraphs or turns in a dialogue. Let such a unit be represented as s_i. In the second stage, these smallest units are treated as single chunks for performing topic-guided retrieval and summarisation via TopicSummRAG (Section 3.2). Overall,

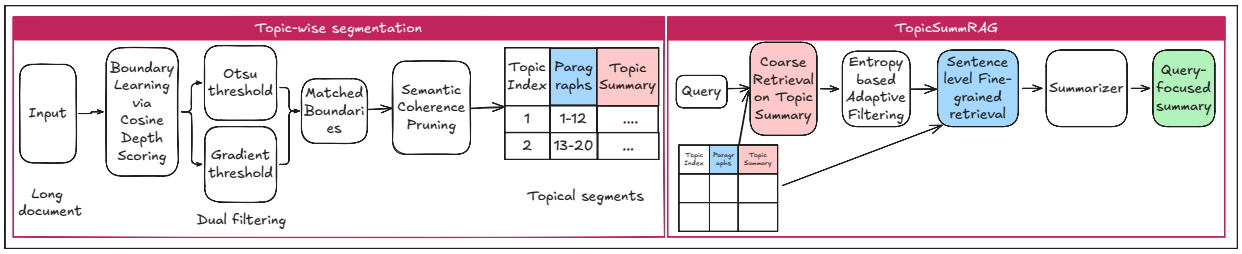

Figure 2: Pipeline of TopicSummRAG

TopicSummRAG performs retrieval and generation in five stages: (1) topical segmentation, (2) topic-aware coarse retrieval using segment-level summaries, (3) entropy-based adaptive selection, (4) sentence-level fine-grained retrieval within selected regions, and (5) final answer generation using a RAG model. A detailed pipeline with example is provided in Appendix A.3.

3.1 Topical segmentation

We conceptualise topics as contiguous discourse units—spans of text that are internally coherent and separated by semantic shifts. When retrieval units correspond to such coherent discourse segments, relevance scoring becomes more stable and retrieved content is more likely to support grounded generation. This insight motivates our decision to treat topical segmentation as a first-class operation, rather than as a downstream heuristic layered on top of retrieval. A central challenge in topical segmentation is that boundary detection is inherently a contextual decision. Topic transitions are better characterised by changes in contextual semantics than by surface cues: while abrupt transitions (e.g., agenda changes in meetings or plot twists in a story) often produce sharp semantic breaks, many documents exhibit gradual topical drift, where relevance accumulates or dissipates slowly across units. Consequently, whether a unit signals topic continuation or topic shift cannot be determined in isolation, but only relative to its surrounding discourse. Two adjacent units may be semantically similar yet belong to different topics, while units within the same topic may exhibit substantial lexical variation. As a result, naive local similarity measures and embeddings optimised solely for content similarity are insufficient for reliable boundary detection. Documents are composed of structural units such as sentences, paragraphs, or dialogue turns; however, their semantic organization is governed by higher-level logical components, including sections, subtopics, and discussion threads. Because these topical boundaries do not correspond to consistent surface-level cues, segmenting text into semantically meaningful spans remains a challenging problem. Our approach addresses this challenge by identifying topical boundaries in a data-driven and domain-general manner.

Learning Contextual Embeddings: Standard text encoders produce representations based on local content similarity, but they are not trained to distinguish semantic continuity from semantic change across discourse. To create embeddings that can account for topical coherence, we introduce a context-aware contrastive pre-training framework. Using `unsup-simcse-roberta-base`[1], each unit s_i (turn, sentence, or paragraph depending on domain) is encoded as $h_i = f_\theta(s_i)$, where θ is the parameter of the network learned via a triplet loss function. For each anchor unit s_i, the positive example s^+ is drawn from the same topical span (i.e., semantically aligned neighbouring units) and the negative example s^- is sampled from a different span, representing a topic shift. The triplet loss is defined as: $\mathcal{L}_{\text{triplet}} = \sum_{(s,s^+,s^-)} \max\big(0,\, m + d(\mathbf{s}, \mathbf{s}^+) - d(\mathbf{s}, \mathbf{s}^-)\big)$, where d is the cosine distance and $m = 0.5$. Triplets are constructed using hard negative mining: the most similar in-span unit serves as the positive. In contrast, semantically close out-of-span units are selected as negatives (Figure 5 in Appendix A.2). This encourages the encoder to separate semantically similar but topically distinct units, facilitating precise boundary detection. Encoding such contextual distinctions is critical not only for accurate boundary detection but also for ensuring that downstream retrieval units align with coherent discourse regions rather than isolated sentences. Appendix A.1 shows the segmentation pipeline .

[1] princeton-nlp/unsup-simcse-roberta-base

Boundary Learning using Cosine Depth Scoring: The contextual embeddings learned in the previous step provide a representation space in which semantic continuity and semantic change are explicitly separated. Leveraging these representations, we detect topical boundaries by measuring the sharpness of contextual semantics change across adjacent regions of the document, rather than between individual units.

Specifically, we identify candidate topic boundaries using cosine-depth scoring, which measures semantic discontinuity between left and right contextual windows. For each unit i (sentence, paragraph, or dialogue turn), we compute left and right contextual embeddings $\mathbf{e}_i^{\text{left}}$ and $\mathbf{e}_i^{\text{right}}$ by pooling augmented unit embeddings within a fixed window of size w. The semantic depth score is defined as $\text{depth}_i = 1 - \frac{\mathbf{e}_i^{\text{left}} \cdot \mathbf{e}_i^{\text{right}}}{\|\mathbf{e}_i^{\text{left}}\| \, \|\mathbf{e}_i^{\text{right}}\|}$. By aggregating context on both sides of a position, cosine depth captures both abrupt semantic breaks and gradual topical drift. However, depth signals alone are insufficient, as long documents naturally contain noise and local fluctuations that can produce spurious peaks.

Empirically, we observe that meaningful topic boundaries tend to be both *globally salient*—standing out relative to the overall depth distribution, and *locally abrupt*, corresponding to sharp semantic changes rather than slow drift. This motivates a dual filtering strategy. First, Otsu thresholding (Otsu, 1979) selects globally significant depth peaks by maximizing between-class variance. Second, gradient-based thresholding ensures that retained candidates correspond to local change points. Because these two signals capture complementary aspects of topic change and may peak at slightly different positions due to windowing effects, we reconcile candidates within a small window before further refinement. A position is retained only if it passes both filters, producing a high-recall set of plausible boundaries.

Semantic Coherence Pruning: Even after dual thresholding, candidate boundaries may remain overly dense or redundant, particularly in documents with smooth topical flow. Enforcing a fixed segmentation granularity is undesirable, as different documents—and even different regions within the same document—exhibit varying structural complexity. We therefore introduce semantic coherence pruning to determine whether introducing a candidate boundary would meaningfully improve segment homogeneity. Rather than relying on external supervision or fixed length constraints, coherence pruning uses embedding-based consistency to decide whether a segment should be split.

For a candidate boundary that splits units $i{:}j$, coherence is defined as $\text{coh}(i,j) = \frac{1}{j-i+1} \sum_{k=i}^{j} \frac{\|h'_k\|_2}{\frac{1}{j-i+1} \sum_{l=i}^{j} \|h'_l\|_2 + \epsilon}$, where h'_k denotes the feature-augmented embedding of unit k, $\|\cdot\|_2$ is the ℓ_2 norm, and ϵ is a small constant for numerical stability. The numerator measures the semantic strength of each unit relative to the segment, while the denominator normalizes by the average embedding magnitude across the segment. This measure captures the internal semantic homogeneity of a segment: high coherence indicates a semantically consistent region that should be preserved, whereas low coherence suggests topical drift and motivates splitting.

We apply coherence pruning sequentially after cosine-depth-based candidate selection to regulate final segmentation granularity. To adapt pruning strength across documents, we compute a document-level homogeneity score defined as the ratio of the depth curve's standard deviation to its mean. Documents with smooth topical flow undergo more aggressive pruning, yielding fewer and broader segments, while documents with frequent sharp transitions are pruned conservatively to preserve finer-grained boundaries. Together, cosine-depth scoring determines where boundaries may occur, while coherence pruning decides which of these candidates are retained, producing segmentations that are precise, semantically coherent, and adaptable across document types.

3.2 Topic guided retrieval and summarization

Having segmented the document into semantically coherent topical spans, the remaining challenge is to efficiently identify which of these spans are relevant to a given query. Direct retrieval over full segments can be sensitive to surface-level variation and inefficient when documents contain many topics. Instead, we perform coarse retrieval over compact segment-level summaries, which serve as high-level semantic abstractions of each topical region. This design allows retrieval to operate at the level of discourse topics rather than individual sentences, improving robustness to lexical noise while preserving thematic relevance. Moreover,

operating over summaries yields a compact relevance distribution that can be analyzed to adapt retrieval breadth based on query ambiguity.

TopicSummRAG is retriever-agnostic: it operates with any model capable of producing vector representations for segment summaries or sentences. The framework requires only a similarity function (e.g., cosine similarity) for ranking units during both coarse- and fine-grained retrieval. For clarity, we describe the method assuming a dense encoder such as SBERT, though the retrieval logic remains unchanged for sparse retrievers (e.g., BM25) or discourse-aware models (e.g., Situated Embeddings). The full set of retrievers and generators evaluated is described in Section 4.2.

Topic-aware Coarse Retrieval using Segment-level Summaries: Using our segmentation method (Section 3.1), the document is divided into topical segments, each representing a semantically coherent discourse region. Each segment is summarized using Qwen3-8B.[2] These summaries provide compact, topic-level representations that abstract away sentence-level variability while preserving dominant semantic content. We embed the query and all segment summaries into a shared semantic space using SBERT and compute cosine similarities. The top-k summaries are retained as high-level topical filters, narrowing the search space to segments thematically aligned with the query.

Following this coarse ranking, we retain the top-L segment summaries as a recall-oriented candidate set. The parameter L defines a fixed upper bound on the number of segments considered for further processing and is chosen to ensure coverage rather than precision. Importantly, L does not determine the final retrieval granularity; it merely bounds the search space prior to adaptive selection. The effective number of retrieved segments is determined in the subsequent stage based on the structure of the relevance distribution. Because adaptive selection depends on relative dominance rather than absolute rank, the same thresholds remain valid as L varies.

Entropy-Based Adaptive Selection: Following topic-aware coarse retrieval, we obtain a ranked list of segment-level summaries with cosine similarity scores relative to the query. However, the sharpness of this similarity distribution varies substantially across queries and documents: some queries are tightly aligned with a single topical segment, while others are inherently broad and span multiple segments. Applying a fixed retrieval cutoff in such cases is brittle and can either discard relevant context or introduce excessive noise. To address this, we analyze the distribution of segment relevance scores using a temperature-scaled softmax. Because cosine similarities lie in a compressed numeric range, directly applying softmax can yield overly flat probability distributions even when meaningful relevance gaps exist. We therefore apply temperature scaling with $\tau = 0.3$ prior to normalization, restoring contrast in the similarity distribution in line with standard practice in contrastive retrieval models. Given top-L similarity scores $\{s_1, \ldots, s_L\}$, we compute: $p_i = \frac{\exp(s_i/\tau)}{\sum_{j=1}^{L} \exp(s_j/\tau)}$. We compute the entropy of this distribution, $H = -\sum_{i=1}^{L} p_i \log p_i$, and its normalized form $H_{\text{norm}} = H/\log L$, which serves as a measure of uncertainty in segment relevance. While entropy captures the overall dispersion of relevance mass, we do not use it directly to determine the number of retrieved segments. Instead, the decisive signal is the dominance of the most relevant segment, quantified as $p_{\max} = \max_i p_i$, which measures how much of the total relevance mass is explained by a single topical region. Based on the dominance of the most relevant segment, measured by the maximum softmax probability $p_{\max}$, we adaptively select the number of retrieved segments k. For focused queries ($p_{\max} \geq 0.6$), a single segment accounts for the majority of the relevance mass, and we retain only the top segment (k=1). For weakly multi-topic queries ($0.4 \leq p_{\max} < 0.6$), one dominant segment and a secondary competitor are present, and we retain the top two segments (k=2). For broad queries ($p_{\max} < 0.4$), relevance is distributed across multiple segments, and we retain all top-L segments to preserve coverage.

Under $\tau = 0.3$, these thresholds correspond to meaningful odds ratios between the top segment and its competitors, distinguishing single-topic, near-single-topic, and genuinely multi-topic queries in a distributional rather than heuristic manner. Entropy thus provides a complementary diagnostic signal reflecting uncertainty, while dominance governs the strength of retrieval pruning. This adaptive selection strategy enables TopicSummRAG to remain precise for focused queries while preserving coverage for broader information needs.

[2]https://huggingface.co/Qwen/Qwen3-8B

Table 1: Statistics of (a) topic segmentation datasets (Left) and (b) query-focused summarization datasets (Right). *Tr = Train, Va = Validation, Te = Test; W = Words, U = Utterances, Doc = Document, Q = Query, L = Length.*

Dataset	# Docs			# Words/Doc			Avg.	
	Tr	Va	Te	Min	Avg	Max	W/Section	U/Doc
TIAGE	286	96	97	109	185.1	264	40.4	15.4
QMSum	162	35	35	1371	9521.4	25529	1593.6	334.7

Dataset	# Docs	# Qs	# Pairs
ODSum-story	1190	635	–
ODSum-meeting	232	436	–
QMSum (All)	232	7.8	279
Dataset	**Avg. Doc L.**	**Avg. Q. L.**	**Avg. Sum. L.**
ODSum-story	808.54	10.79	273.80
ODSum-meeting	7176.21	32.89	185.17
QMSum (All)	9069.8	–	69.6

Sentence-level fine-grained retrieval within matched regions: Once relevant segments are identified, we perform second-stage sentence-level retrieval to extract the most evidence-rich text. Each selected segment is decomposed into sentences, forming a fine-grained retrieval corpus $S = \{s_1, s_2, \ldots, s_L\}$. By default, we use SBERT as our dense retriever (we also evaluate BM25 and Situated Embeddings in Section 4.2). The query embedding $\mathbf{q}$ is compared against all sentence embeddings using cosine similarity. We rank sentences and select the top-k_s evidence sentences (we use $k_s = 10$). This produces a targeted evidence set: $E = \{e_1, e_2, \ldots, e_{k_s}\}$ where each e_i is a highly relevant sentence grounded in both local topic and global query semantics.

Topic-level summaries filter which regions matter. Sentence-level retrieval identifies what exact information is relevant inside those regions. This two-tier design balances high recall (via segment selection) with high precision (via sentence retrieval), which was not possible with chunk-level RAG alone.

Final answer generation: We concatenate the retrieved evidence sentences with the query and provide them to a summarizer model for final answer generation (prompt added in Appendix A.7). The model is instructed to generate an answer grounded strictly in retrieved evidence, mitigating hallucination and over-retrieval. The final evidence set thus contains the best matching topical regions and the most relevant sentences within those regions. Together, these ensure that the generator conditions its output on coherent, query-focused, and contextually grounded content.

4 Experiments and Results

We present results on two sets of experiments - Experiment 1 shows our results on the topic segmentation model and Experiment 2 on query-focused summarization.

4.1 Experiment 1 - topic segmentation

Datasets: We evaluate our topic segmentation model on two distinct benchmarks that capture different aspects of conversational dynamics: open-domain social chitchat (TIAGE (Xie et al., 2021)) and multi-party formal meetings (QMSum (Zhong et al., 2021)). The detailed statistics for both datasets are summarised in Table 1 and further details are added in Appendix A.4.1.

Baselines: To evaluate the effectiveness of our proposed unsupervised segmentation approach, we compare against a comprehensive set of baseline models - SumSeg (Artemiev et al., 2024), MiniSeg (Retkowski & Waibel, 2024), DialogLM (Zhong et al., 2022), BiLSTM_RoBERTa (Ghinassi et al., 2023), TopSeg_RoBERTa (Ghinassi et al., 2023), Deep Tiling with Sentence Embeddings (`all-MiniLM-L6-v2`[3] and `all-mpnet-base-v2`[4]) and Gemini [5]. Further details are added in Appendix A.4.2.

Metrics: To assess the performance of our segmentation model, we employ a combination of boundary-based classification metrics (Precision, Recall, and the F1-score) and distance-based structural metrics (P_k metric (Beeferman et al., 1999) and WindowDiff (Pevzner & Hearst, 2002)). This dual evaluation captures

[3]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[4]https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[5]gemini-1.5-flash

Table 2: Results of topic segmentation models on QMSum (Left) and TIAGE (Right) datasets. * indicates results reported directly from the respective papers. *Note:* F1 is computed as the macro-average of per-document F1 scores

Model	P	R	F1	P_k	WD
SumSeg*	–	–	–	0.357	0.379
MiniSeg*	**0.3109**	0.1692	0.2192	–	–
DialogLM*	–	–	–	0.380	0.400
BiLSTM	0.0644	0.1406	0.0313	0.4121	0.4830
TopSeg	0.0134	0.0305	0.0181	0.3531	0.3830
MiniLM	0.0290	0.0240	0.0260	0.4580	0.4830
Mpnet	0.0310	0.0280	0.0300	0.4370	0.4660
Gemini	0.0740	0.3600	0.1170	0.2640	0.3070
Ours	0.1716	**0.5631**	**0.2630**	**0.3579**	**0.3789**

Model	P	R	F1	P_k	WD
SumSeg*	–	–	–	0.438	0.455
BiLSTM	0.0580	0.1280	0.0790	0.440	0.500
TopSeg	0.0120	0.0270	0.0170	0.390	0.410
MiniLM	0.0260	0.0220	0.0240	0.480	0.505
Mpnet	0.0280	0.0250	0.0260	0.460	0.490
Gemini	0.0680	0.3300	0.1120	0.300	0.340
Ours	**0.1600**	**0.5300**	**0.2450**	**0.370**	**0.390**

both boundary-level accuracy and global structural consistency, which is essential for assessing segmentation quality in long, multi-topic documents. Further details are added in Appendix A.4.3.

Results: Table 2 reports segmentation results on both the QMSum and TIAGE datasets. On QMSum, our approach achieves the highest F1 score (0.263) and the highest recall (0.5631), indicating that it recovers a larger proportion of true segment boundaries while maintaining substantially lower boundary error rates.. Although MiniSeg attains the highest precision (0.3109), its recall is substantially lower (0.1692), suggesting a conservative segmentation strategy that misses many true boundaries. Traditional baselines such as BiLSTM and TopSeg perform poorly across all metrics, with low F1 scores and higher P_k and WD values, reflecting limited ability to capture topic transitions in long, multi-turn meetings. Pretrained embedding-based models (MiniLM and Mpnet) show modest improvements over these baselines but remain clearly behind our method, particularly in recall. Gemini achieves relatively high recall (0.360) but low precision (0.074), leading to over-segmentation and only moderate overall performance.

On the TIAGE dataset, similar trends are observed, though overall performance is lower across methods. This reflects the lack of explicit discourse structure and frequent topical drift in open-domain social dialogue. Our model again achieves the best F1 score (0.245) and recall (0.530), demonstrating its ability to generalize beyond meeting transcripts. Baseline methods such as BiLSTM, TopSeg, MiniLM, and Mpnet exhibit low precision and F1 scores, indicating difficulty in reliably identifying segment boundaries. Gemini continues to favor recall (0.330) at the expense of precision (0.068), resulting in over-segmentation. Importantly, our approach also attains the lowest P_k (0.370) and WD (0.390) values on TIAGE, suggesting more accurate boundary placement compared to other methods. These results show that unsupervised, recall-oriented boundary detection with controlled structural error effectively supports topic-guided retrieval.

Error analysis: Error patterns differ across datasets and reflect their underlying discourse structure. On QMSum, errors often occur during gradual topical drift, for example when a meeting transitions from an agenda recap into a decision-making phase without an explicit shift in vocabulary; such transitions may not yield sharp cosine-depth peaks, leading to occasional under-segmentation. On TIAGE, errors are more frequently driven by rapid speaker alternation and informal conversational style, where lexical variation across turns can trigger spurious boundaries despite stable topical content. While semantic coherence pruning mitigates many of these cases, distinguishing discourse-level topic change from surface-level conversational dynamics remains challenging.

4.2 Experiment 2 - Query focused summarization

Datasets: We evaluate the performance of our topic-guided query-focused summarisation model across three distinct benchmarks, covering narrative storytelling (ODSum-Story (Rao et al., 2024)) to complex, multi-party meeting transcripts (QMSum and ODSum-Meeting (Rao et al., 2024)), all of which provide query-conditioned reference summaries. Further details are added in Appendix A.5.1.

Baselines: To evaluate the performance of query-focused summarisation models, we conducted our experiments using various retrieval and answer generation models. For retrieval, we experiment with three different

algorithms: BM25 (Robertson & Zaragoza, 2009), Sentence-BERT (SBERT) (Reimers & Gurevych, 2019), and Situated Embeddings (Situated) (Wang et al., 2023). Further details are added in Appendix A.5.2.

For answer generation, we evaluated three state-of-the-art open-source LLMs: LLaMA-3.1-8B[6], Gemma-3[7], and Qwen-3-8B, enabling assessment of our method's robustness across diverse model architectures. Following prior work on chunking mechanisms used in RAG (Gao et al., 2024), we compared with three different RAG chunking mechanisms: (1) Fixed-size, splitting documents into contiguous spans of predefined length; (2) Sentence, treating each sentence as an independent retrieval unit; and (3) Semantic, segmenting documents based on embedding similarity between adjacent sentences using an off-the-shelf library.

Metrics: Inspired by Kirstein et al. (2024), we evaluate summary quality using a complementary set of metrics that capture lexical overlap (ROUGE-1,2,L; (Lin, 2004)), semantic alignment (LENS; (Maddela et al., 2023)), and referenceless functional utility (BLANC; (Vasilyev et al., 2020)). Together, these metrics provide a holistic assessment of model performance, accounting for diverse error types such as omissions, hallucinations, and coherence issues. Additional details are provided in Appendix A.5.3.

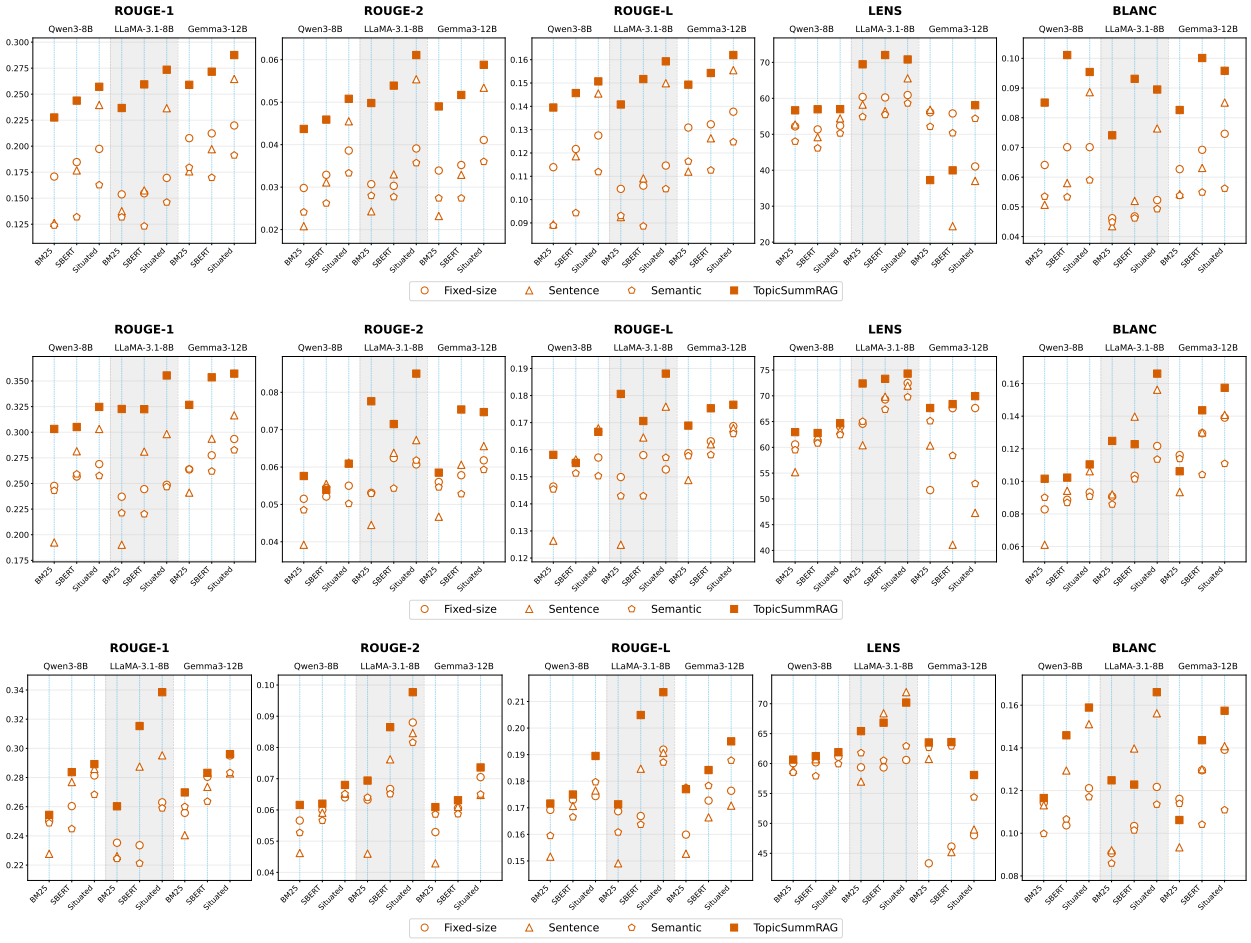

Figure 3: Results on ODSum-Story (top), ODSum-Meeting (middle), and QMSum (bottom) datasets.

Results: Figures 3 compare TopicSummRAG against fixed-size, sentence-level, and semantic chunking across ODSum-Story, ODSum-Meeting, and QMSum under multiple retrieval strategies. Across all datasets, TopicSummRAG consistently improves query-focused summarization quality, indicating that aligning retrieval units with topical structure yields more relevant and coherent evidence for generation. On ODSum-Story and ODSum-Meeting, TopicSummRAG achieves substantial gains over all baselines across ROUGE,

[6]meta-llama/Llama-3.1-8B-Instruct
[7]google/gemma-3-12b-it

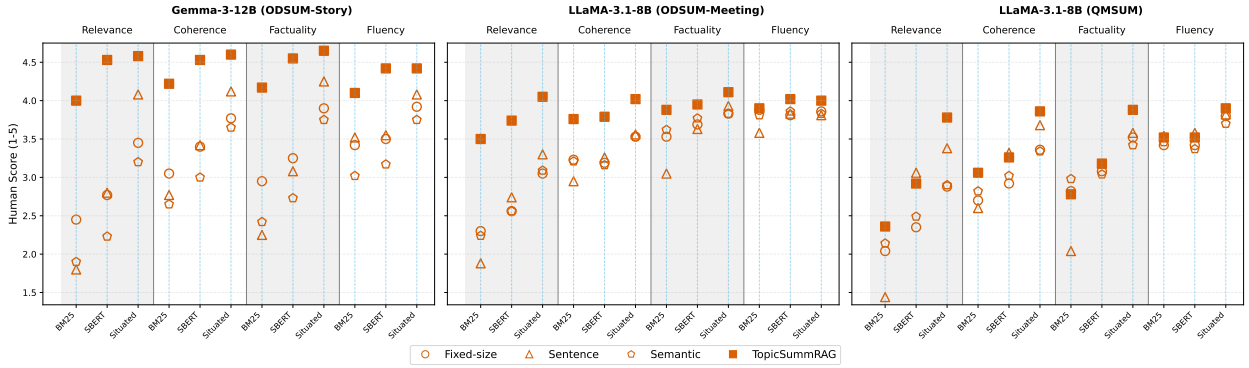

Figure 4: G-EVAL scores

LENS, and BLANC metrics. In particular, improvements in ROUGE-2 and ROUGE-L indicate better coverage of query-relevant content and improved sequence-level coherence, while higher LENS and BLANC scores suggest that retrieved segments form semantically unified and generation-friendly contexts. These gains are consistent across BM25, SBERT, and Situated retrieval, demonstrating that TopicSummRAG complements both lexical and semantic retrievers. On QMSum, TopicSummRAG similarly outperforms baseline chunking strategies across all evaluated generators. Gains in ROUGE-2 and ROUGE-L reflect improved grounding in query-relevant discourse spans, while higher LENS and BLANC scores confirm increased semantic coherence and factual utility. Importantly, improvements persist across different generator architectures, indicating that the benefits of topic-aware chunking are model-agnostic rather than tied to a specific LLM.

Table 4 reports the G-EVAL scores on top-performing models from each dataset such as ODSum-Story (Gemma-3-12B), ODSum-Meeting (LLaMA-3.1-8B), and QMSum (LLaMA-3.1-8B) (details in Appendix A.8). TopicSummRAG consistently achieves the consistently competitive or best-performing scores across all metrics and retrieval methods, demonstrating that topic-aware chunking produces semantically coherent, informative, and fluent summaries. High G-EVAL scores indicate that summaries are factually accurate, contextually coherent, topically aligned, and easily interpretable by downstream models. On the ODSum datasets, it yields notable gains in Relevance, Coherence, and Factuality, while on QMSum it similarly outperforms baseline chunking strategies, confirming its robustness and generalizability across models and meeting domains. Overall, these results confirm that topic-aware segmentation improves query-focused summarization by providing retrieval units that are both semantically coherent and aligned with the underlying discourse structure, across datasets, retrievers, and generators.

Error analysis: We analyse cases where TopicSummRAG underperforms baseline chunking on QMSum. For highly specific queries targeting a single action item or decision, topic-level chunks may include additional surrounding discussion that is topically related but not strictly necessary to answer the query, reducing relevance under dense retrieval. Under BM25 retrieval, summarizing multiple related utterances within a topic can occasionally lead to mild over-generalization, where fine-grained details are merged, which is penalized by G-EVAL factuality scores. Notably, these errors do not affect fluency or coherence, indicating that failures stem from context aggregation rather than hallucination. A more detailed error analysis and significant t-test is added in Appendix A.10 and A.11.

Ablation study We conduct a systematic ablation study to isolate the contribution of individual components in both the topical segmentation model and the TopicSummRAG retrieval pipeline. By selectively removing or modifying key modules while holding all other factors constant, we evaluate how each design choice affects boundary detection quality and downstream query-focused retrieval performance. This analysis aims to validate that the observed gains stem from principled alignment between topical structure and retrieval behaviour, rather than from increased model capacity or incidental design choices.

Table 3: Combined ablation studies: (Left) Segmentation pipeline; (Right) TopicSummRAG results on QMSum. (E: Entropy, *summ*: summarization).

Method	P (↑)	R (↑)	F1 (↑)	P_k (↓)	WD (↓)
Ours	**0.17**	0.56	**0.26**	0.36	0.38
w/o contrastive	0.12	0.56	0.14	0.39	0.38
w/o Okta	0.08	**0.59**	0.13	**0.32**	**0.33**
w/o Grad	0.10	0.55	0.15	0.38	0.40
w/o Okta & Grad	0.14	0.71	0.21	0.37	0.39
w/o semantic pruning	0.11	0.44	0.15	0.36	0.38

Method	P (↑)	R (↑)	F1 (↑)
Sentence RAG	0.17	0.36	0.21
TopicSumm	0.02	**0.86**	0.04
TopicSumm + E	0.07	0.78	0.12
TopicSumm + E → RAG	**0.39**	0.26	**0.27**
w/o chunk-level summ	0.28	0.25	0.26

Topic Segmentation pipeline ablation: Table 3 (Left) summarises the contribution of each module to our segmentation pipeline. Precision here reflects boundary reliability, i.e., the proportion of predicted boundaries that align with true topic transitions. The complete model achieves the highest Precision (0.17) and F1 score (0.26), suggesting that each stage of the architecture contributes meaningfully to overall performance. The most significant performance degradation occurs when contrastive pre-training is removed, resulting in a drop in the F1 score from 0.26 to 0.14. This confirms that pre-training is essential for generating the discriminative embeddings required to identify topic transitions. Similarly, omitting semantic pruning results in a 42% decrease in F1 (0.15), indicating that local boundary detection alone is insufficient without a final coherence check to remove spurious segments. The Okta filtering module primarily governs the model's Precision. Removing Okta results in a sharp decline in Precision (from 0.17 to 0.08) and a corresponding increase in Recall. Although the 'w/o Okta' variant yields lower P_k and WD scores, this reflects a known artefact of distance-based metrics under over-segmentation, where frequent boundary predictions reduce average boundary distance without improving boundary correctness. The substantial drop in F1 confirms that Okta filtering is essential for maintaining boundary reliability. Finally, the Gradient filtering module serves as a secondary noise-reduction step. Removing it increases P_k and WD errors by 0.02, signalling a loss in boundary placement accuracy. When both Okta and Gradient filters are disabled, the system defaults to an aggressive segmentation strategy (Recall: 0.71), which fails to maintain the structural integrity of the topics. Overall, the results support a multi-stage segmentation pipeline with contextual pre-training, boundary filtering, and semantic pruning.

TopicSummRAG pipeline ablation: Table 3 (Right) evaluates different retrieval configurations. The Sentence RAG baseline achieves moderate F1 (0.218) but suffers from noisy retrieval, including many irrelevant sentences. The TopicSumm module alone yields a very high recall (0.866), effectively capturing relevant topical chunks for the query. Introducing entropy-based adaptive selection further reduces retrieval noise by pruning weakly relevant topical segments, improving precision while preserving relevant content. Combining TopicSumm, entropy filtering, and RAG retrieval produces the best-balanced performance, achieving the highest precision (0.392) and F1 score (0.272). This progression demonstrates how each stage incrementally cleans the retrieval output, from broad coverage to focused, concise, and query-relevant results. Removing chunk-level summarisation forces coarse retrieval to operate over raw segment embeddings, which often mix multiple subtopics and dilute topical signals, leading to noisier retrieval and reduced precision. Although recall remains comparable due to broader coverage, the resulting context is noisier, leading to reduced precision and F1. This result highlights that topic segmentation alone is insufficient: high recall must be complemented by adaptive selection and fine-grained retrieval to achieve precise, query-focused evidence.

The ablations demonstrate that TopicSummRAG's gains arise from the joint interaction of contextual segmentation, adaptive filtering, and fine-grained retrieval, with each component contributing to either structural coherence or retrieval precision. To further assess robustness beyond component-level ablations, we report additional sensitivity analyses in Appendix A.6, examining the effects of retrieval granularity, temperature scaling, contextual window size, and candidate set size; these results confirm that TopicSummRAG's gains are stable across a broad range of hyperparameter choices.

Hyperparameters We finetune the `unsup-simcse-roberta-base` encoder using the proposed context-aware contrastive learning objective and generate unit-level embeddings from the finetuned model. Embeddings are computed with a maximum token length of 256 and chunked inference (batch size 50), ensuring stable GPU utilisation across documents of varying lengths. For topical boundary detection, semantic depth is computed using a sliding contextual window of size W. We set $W = 3$ for short dialogues and $W = 5$

for longer documents, reflecting a trade-off between sensitivity to local semantic transitions and robustness to conversational noise. To avoid repeated detections of the same topical transition, we enforce a minimum inter-peak distance of 3 units for short documents and 7 units for longer meetings, which prevents over-fragmentation while preserving genuine boundary candidates. We incorporate lightweight structural cues via a feature augmentation module consisting of a projection layer (32 or 64 dimensions) followed by a transformer mixer with 1–2 layers and 2–4 attention heads. This design introduces limited contextual interaction to stabilise depth estimates without adding substantial computational overhead. Larger configurations did not yield measurable improvements in preliminary experiments. Candidate boundary selection relies on two complementary detectors: Otsu thresholding, which provides a global, distribution-based cutoff over depth values, and gradient-based detection, which identifies sharp local semantic changes. A boundary is retained only if both detectors agree within a ± 3-unit window, accommodating minor positional offsets introduced by contextual windowing. Following candidate selection, semantic coherence pruning regulates final segmentation granularity. Boundaries whose resulting segments exhibit cross-segment coherence above 0.99 are removed, as such values indicate near-identical semantic distributions across the split. A minimum segment length of three units is enforced to eliminate trivial partitions.

For topic-guided retrieval, each segment is summarised using Qwen3-8B and query–segment similarities are computed using SBERT. To correct for the compressed numeric range of cosine similarity scores, we apply temperature scaling with $\tau = 0.3$ before softmax normalisation, following standard practice in contrastive retrieval models. We set $L = 5$ for QMSum and $L = 10$ for ODSum, reflecting differences in topical breadth across datasets; sensitivity to L is analysed in Appendix A.6. The number of retained segments k is determined adaptively based on the dominance of the relevance distribution: $k = 1$ when the maximum softmax probability exceeds 0.6, $k = 2$ when it lies between 0.4 and 0.6, and $k = L$ otherwise. Entropy is computed to characterise uncertainty in the similarity distribution but is not used directly to determine k. In the fine-grained retrieval stage, selected segments are decomposed into sentences, and the top $k_s = 10$ sentences are retrieved using dense similarity matching. This value balances evidence coverage with prompt length constraints and is held fixed across datasets.

5 Conclusion

We introduced TopicSummRAG, a framework that addresses the limitations of naïve chunking in retrieval-augmented generation for long, multi-topic documents. The approach is built on a topic segmentation pipeline that partitions documents into semantically coherent discourse units, aligning retrieval with latent topical structure rather than arbitrary textual boundaries. Evaluations on QMSum and TIAGE show that this segmentation model achieves more accurate boundary detection than strong baselines, particularly for long, multi-turn conversations. Building on these segments, TopicSummRAG applies segment-level summarization and dominance-based adaptive retrieval to regulate retrieval granularity. This design improves retrieval precision for focused queries while preserving coverage for broader information needs. Across multiple retrievers and LLMs, TopicSummRAG consistently improves factual grounding, coherence, and summary quality. These results highlight the importance of topic-aware indexing for robust long-document RAG, though highly localized queries may occasionally benefit from finer-grained retrieval units.

6 Limitation

TopicSummRAG introduces additional computation due to segment-level summary generation prior to retrieval, although this overhead is lower than approaches that rely on full LLM-based segmentation or repeated long-context inference (Lee et al., 2025; Duarte et al., 2024; McCarthy et al., 2023). We do not provide a detailed quantitative comparison of preprocessing cost or end-to-end latency, which we leave for future work. Additionally, because retrieval operates over coherent topical regions, highly localized queries may occasionally include limited surrounding context, reflecting a trade-off between discourse coherence and minimal evidence selection.

References

Aleksei Artemiev, Daniil Parinov, Alexey Grishanov, Ivan Borisov, Alexey Vasilev, Daniil Muravetskii, Aleksey Rezvykh, Aleksei Goncharov, and Andrey Savchenko. Leveraging summarization for unsupervised dialogue topic segmentation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4697–4704, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.291. URL https://aclanthology.org/2024.findings-naacl.291/.

Pinkesh Badjatiya et al. Evaluating text segmentation using boundary edit distance. 2018. https://www.researchgate.net/publication/237073700_Evaluating_Text_Segmentation_using_Boundary_Edit_Distance.

Tal Baumel, Raphael Cohen, and Michael Elhadad. Topic concentration in query focused summarization datasets. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 2573–2579. AAAI Press, 2016.

Tal Baumel, Matan Eyal, and Michael Elhadad. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models, 2018. URL https://arxiv.org/abs/1801.07704.

Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Mach. Learn.*, 34(1–3):177–210, February 1999. ISSN 0885-6125. doi: 10.1023/A:1007506220214. URL https://doi.org/10.1023/A:1007506220214.

Andrew Brown, Muhammad Roman, and Barry Devereux. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *Big Data and Cognitive Computing*, 9(12), 2025. ISSN 2504-2289. doi: 10.3390/bdcc9120320. URL https://www.mdpi.com/2504-2289/9/12/320.

Jack G. Conrad, Jochen L. Leidner, Frank Schilder, and Ravi Kondadadi. Query-based opinion summarization for legal blog entries. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pp. 167–176, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585970. doi: 10.1145/1568234.1568253. URL https://doi.org/10.1145/1568234.1568253.

André V. Duarte, João DS Marques, Miguel Graça, Miguel Freire, Lei Li, and Arlindo L. Oliveira. LumberChunker: Long-form narrative document segmentation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6473–6486, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.377. URL https://aclanthology.org/2024.findings-emnlp.377/.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025. URL https://arxiv.org/abs/2404.16130.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Iacopo Ghinassi. Unsupervised text segmentation via deep sentence encoders: a first step towards a common framework for text-based segmentation, summarization and indexing of media content. unsupervised text segmentation via deep sentence encoders. 05 2021. doi: 10.5281/zenodo.4744399.

Irene Ghinassi et al. Comparing neural sentence encoders for topic segmentation across domains: not your typical text similarity task. *PeerJ Computer Science*, 9:e1593, 2023. doi: 10.7717/peerj-cs.1593. URL https://peerj.com/articles/cs-1593/.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. Late chunking: Contextual chunk embeddings using long-context embedding models, 2025. URL https://arxiv.org/abs/2409.04701.

Marti A Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64, 1997.

Yubin Hong, Chaofan Li, Jingyi Zhang, and Yingxia Shao. Fg-rag: Enhancing query-focused summarization with context-aware fine-grained graph rag, 2025. URL https://arxiv.org/abs/2504.07103.

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. What's under the hood: Investigating automatic metrics on meeting summarization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6709–6723, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.393. URL https://aclanthology.org/2024.findings-emnlp.393/.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. Text segmentation as a supervised learning task. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 469–473, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2075. URL https://aclanthology.org/N18-2075/.

Seungmin Lee, Yongsang Yoo, Minhwa Jung, and Min Song. Def-DTS: Deductive reasoning for open-domain dialogue topic segmentation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20736–20753, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1066. URL https://aclanthology.org/2025.findings-acl.1066/.

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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. LENS: A learnable evaluation metric for text simplification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16383–16408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.905. URL https://aclanthology.org/2023.acl-long.905/.

Arya McCarthy, Hao Zhang, Shankar Kumar, Felix Stahlberg, and Ke Wu. Long-form speech translation through segmentation with finite-state decoding constraints on large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 247–257, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.19. URL https://aclanthology.org/2023.findings-emnlp.19/.

Inderjeet Nair, Aparna Garimella, Balaji Vasan Srinivasan, Natwar Modani, Niyati Chhaya, Srikrishna Karanam, and Sumit Shekhar. A neural CRF-based hierarchical approach for linear text segmentation. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 883–893, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.65. URL https://aclanthology.org/2023.findings-eacl.65/.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002. doi: 10.1162/089120102317341756. URL https://aclanthology.org/J02-1002/.

Jinfeng Rao, Yichong Xu, Chenguang Xu, Yiming Yang, and Liwei Wang. Odsum: Query-focused summarization of open-domain conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Fabian Retkowski and Alexander Waibel. From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 406–419, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.25. URL https://aclanthology.org/2024.eacl-long.25/.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Jainisha Sankhavara and Prasenjit Majumder. Query-focused biomedical text summarization in bioasq 8b. In *CLEF (Working Notes)*, 2020.

Haiyue Song, Raj Dabre, Zhuoyuan Mao, Chenhui Chu, and Sadao Kurohashi. BERTSeg: BERT based unsupervised subword segmentation for neural machine translation. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 85–94, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-short.12. URL https://aclanthology.org/2022.aacl-short.12/.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. Fill in the BLANC: Human-free quality estimation of document summaries. In Steffen Eger, Yang Gao, Maxime Peyrard, Wei Zhao, and Eduard Hovy (eds.), *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 11–20, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.eval4nlp-1.2. URL https://aclanthology.org/2020.eval4nlp-1.2/.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. SQuALITY: Building a long-document summarization dataset the hard way. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1139–1156, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.75. URL https://aclanthology.org/2022.emnlp-main.75/.

Yile Wang, Yifan Gao, Zhiwei Li, Lidong Bing, and Wai Lam. Situated summarization: Leveraging discourse context for query-focused long-document summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Zhitong Wang, Cheng Gao, Chaojun Xiao, Yufei Huang, Shuzheng Si, Kangyang Luo, Yuzhuo Bai, Wenhao Li, Tangjian Duan, Chuancheng Lv, Guoshan Lu, Gang Chen, Fanchao Qi, and Maosong Sun. Document segmentation matters for retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8063–8075, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.422. URL https://aclanthology.org/2025.findings-acl.422/.

Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. TIAGE: A benchmark for topic-shift aware dialog modeling. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1684–1690, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.145. URL https://aclanthology.org/2021.findings-emnlp.145/.

Eric Yang, Jonathan Amar, Jong Ha Lee, Bhawesh Kumar, and Yugang Jia. The geometry of queries: Query-based innovations in retrieval-augmented generation for healthcare qa, 2025. URL https://arxiv.org/abs/2407.18044.

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. Financial report chunking for effective retrieval augmented generation, 2024. URL https://arxiv.org/abs/2402.05131.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205/.

Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. Meta-chunking: Learning text segmentation and semantic completion via logical perception, 2025. URL https://arxiv.org/abs/2410.12788.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. Qmsum: A new benchmark for query-based multi-domain meeting summarization, 2021. URL https://arxiv.org/abs/2104.05938.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Dialoglm: Pre-trained model for long dialogue understanding and summarization, 2022. URL https://arxiv.org/abs/2109.02492.

A Appendix

A.1 Algorithm for Contextually Coherent Topic Segmentation (CCDS)

Algorithm 1 Contextually Coherent Topic Segmentation (CCDS)

Require: Transcript $T = \{t_1, \ldots, t_N\}$, dialogue embedder $\mathcal{E}$, window size W, minimum boundary index m_b, feature augmenter $\mathcal{A}$, hyperparameters Θ

Ensure: Predicted topic boundaries $\mathcal{B}$

$E \leftarrow [\, e_i = \mathcal{E}(t_i)\,]_{i=1}^{N}$ ▷ contextual embeddings

$E' \leftarrow \mathcal{A}(E)$ ▷ feature augmentation

for $i \leftarrow 1$ **to** N **do**

  $L_i \leftarrow \text{mean}\Big(E'_{\max(1,i-W):i-1}\Big)$ ▷ left context

  $R_i \leftarrow \text{mean}\Big(E'_{i+1:\min(N,i+W)}\Big)$ ▷ right context

  $d_i \leftarrow \text{cosine_depth}(e_i, L_i, R_i)$

end for

$D \leftarrow \{d_1, \ldots, d_N\}$ ▷ depth signal

peaks $\leftarrow$ argrelextrema$(D, \text{order} = \lfloor W/2 \rfloor)$

$\tau \leftarrow \text{OtsuThreshold}(D)$

$\mathcal{C}_1 \leftarrow \{p \in \text{peaks} \mid d_p \geq \tau\}$

$\mathcal{G} \leftarrow \text{gradient_candidates}(D)$

$\mathcal{C}_2 \leftarrow \{p \in \mathcal{C}_1 \mid \exists g \in \mathcal{G}, |p - g| \leq W\}$

Compute homogeneity $H \leftarrow \sigma(D)/(\mu(D) + \epsilon)$

Adapt pruning thresholds in Θ using H

$\mathcal{B} \leftarrow \pi(\mathcal{C}_2, D, E; \Theta)$ ▷ semantic coherence pruning

$\mathcal{B} \leftarrow \{b \in \mathcal{B} \mid b \geq m_b\}$

Ensure boundary at index $N - 1$

return sorted unique $\mathcal{B}$

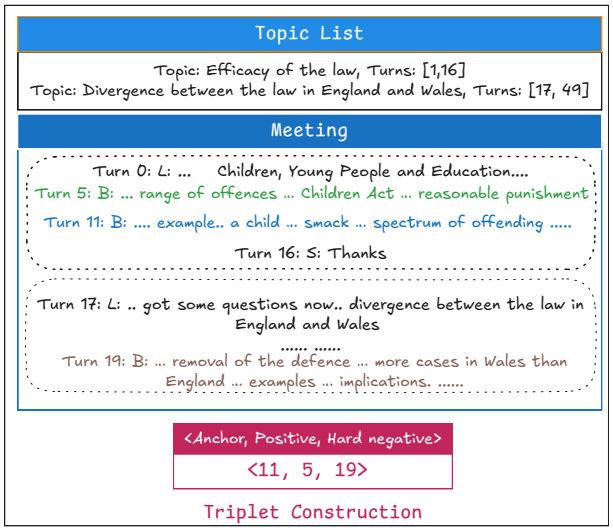

Figure 5: Topic 1 (Turns 1–16) covers efficacy of the law; Topic 2 (Turns 17–49) covers divergence between England and Wales. Turn 11 (Anchor) is paired with Turn 5 (Positive, same topic) and Turn 19 (Hard Negative, different topic but semantically similar). Gold topic spans enable mining of positives and hard negatives that promote embeddings separating semantically overlapping yet topically distinct turns.

A.2 Example of context-aware triplet construction on QMSum meeting

Figure 5 shows the example of creating triplets for contrastive model training.

A.3 Detailed pipeline

Figure 6 shows the detailed step-by-step topic segmentation pipeline and TopicSummRAG using an example of a long story.

A.4 Experiment 1- Topic Segmentation

A.4.1 Datasets

We evaluate our topic segmentation model on two distinct benchmarks that capture different aspects of conversational dynamics: open-domain social chitchat (TIAGE (Xie et al., 2021)) and multi-party formal meetings (QMSum (Zhong et al., 2021)).

TIAGE (Xie et al., 2021) is a specialized benchmark constructed by augmenting the PersonaChat dataset (Zhang et al., 2018) with human-annotated topic shifts. Unlike general dialogue datasets, TIAGE focuses explicitly on the transition points where a conversation fluently moves from one persona-based topic to another. The dataset contains 7,861 gold-standard topic-shift annotations across 500 multi-turn dialogues. We utilise the human-annotated portion of the dataset, which provides a high-quality ground truth for evaluating a model's ability to detect natural conversational flow and proactive topic transitions. QMSum (Zhong et al., 2021) is a large-scale, multi-domain benchmark originally designed for query-based meeting summarisation. It consists of 232 meetings with 1,808 query-summary pairs across three domains: Product (AMI), Academic (ICSI), and Committee discussions. The detailed statistics for both datasets are summarised in Table **??**. Although both datasets are conversational, they differ substantially in discourse structure and topical granularity, allowing us to evaluate robustness across both informal and highly structured settings.

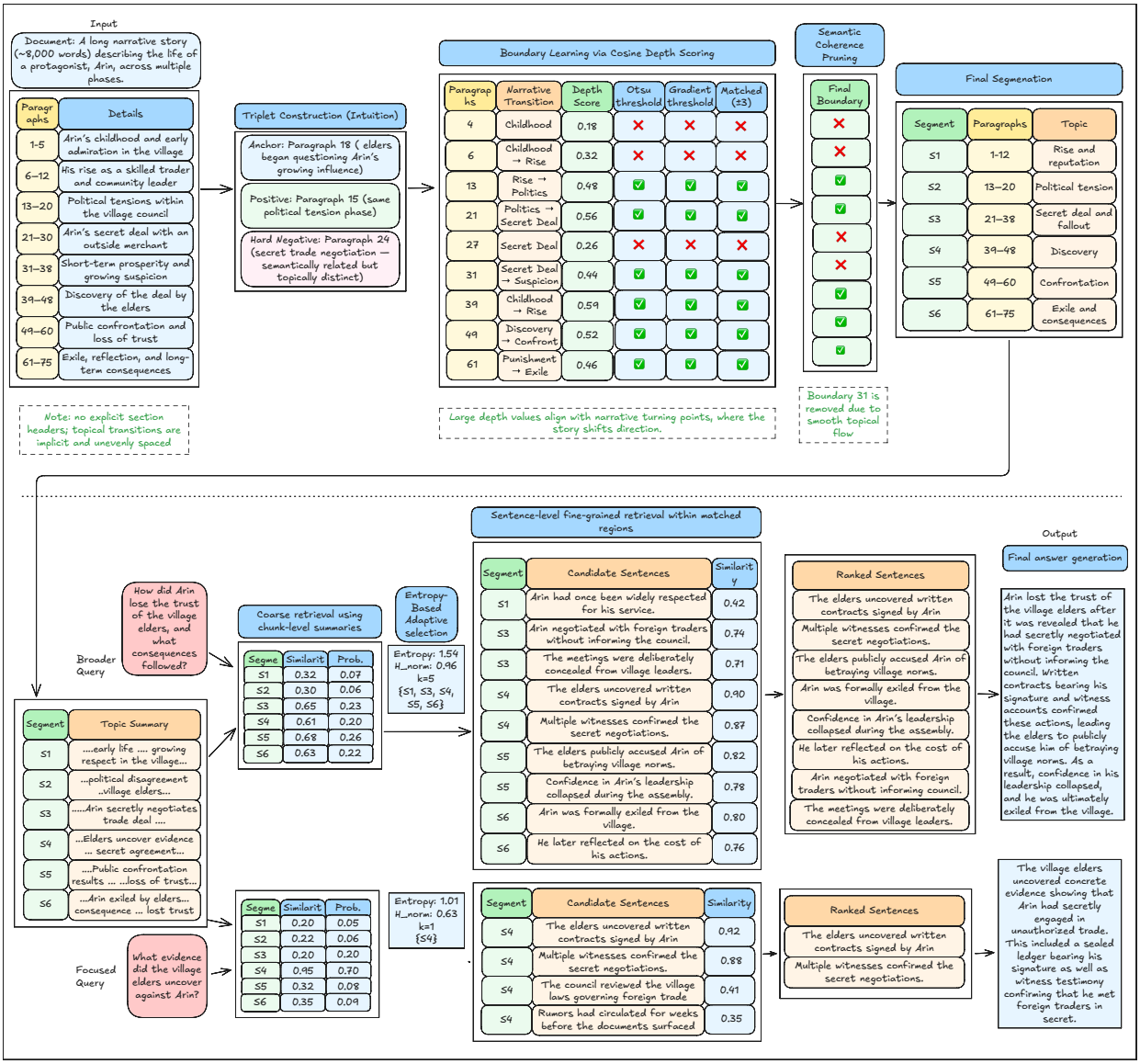

Figure 6: Pipeline of TopicSummRAG (below) and Topic Segmentation (above)

A.4.2 Baselines for Topic Segmentation model

To evaluate the effectiveness of our proposed unsupervised segmentation approach, we compare against a comprehensive set of baseline models listed below.

1. SumSeg (Artemiev et al., 2024): An unsupervised dialogue topic segmentation method that leverages summarization signals for boundary detection.

2. MiniSeg (Retkowski & Waibel, 2024): A hierarchical text segmentation model originally designed for smart chaptering of unstructured spoken content

3. DialogLM (Zhong et al., 2022): A dialogue-specific pretrained language model designed for meeting summarization and segmentation tasks.

4. BiLSTM_RoBERTa (Ghinassi et al., 2023): A hybrid encoder that combines RoBERTa embeddings with a bidirectional LSTM layer to capture both contextual token representations and sequential discourse dependencies.

5. TopSeg_RoBERTa (Ghinassi et al., 2023): TopSeg applies the DeepTiling (Ghinassi, 2021) algorithm on sentence embeddings generated by RoBERTa.

6. Deep Tiling with Sentence Embeddings: In addition to reported baselines, we apply the deep tiling segmentation technique using two widely adopted sentence-transformer models `all-MiniLM-L6-v2`[8] and `all-mpnet-base-v2`[9].

7. Gemini: We used `gemini-1.5-flash` in a zero-shot setting, where segmentation is obtained by directly prompting the model.

For SumSeg, MiniSeg and DialogLM, we report segmentation results directly from the respective papers to ensure faithful comparison with published benchmarks. For other baselines, we have conducted experiments locally under consistent settings using publicly available code and models.

A.4.3 Metrics

Classification Metrics The standard metrics include Precision, Recall, and the F1-score. In the context of text segmentation, these are calculated by treating each potential boundary position (e.g., sentence boundary) as a binary classification point. Precision denotes the proportion of predicted boundaries that align with the ground truth; Recall measures the proportion of actual boundaries correctly identified by the model; and the F1-score serves as the harmonic mean of the two to provide a balanced assessment of overall accuracy.

Distance-Based Metrics Standard classification metrics are often insufficient for segmentation because they do not account for "near misses" (e.g., a boundary placed one sentence away from the truth). To address this, we utilize P_k and WindowDiff, which are the de facto standards in the topic segmentation field (Retkowski & Waibel, 2024; Badjatiya et al., 2018). The P_k metric (Beeferman et al., 1999) measures the probability that two sentences chosen at random at a distance k are incorrectly classified as being in the same or different segments. A lower P_k score indicates higher segmentation quality. WindowDiff (Pevzner & Hearst, 2002) improves upon P_k by moving a sliding window of size k across the text and comparing the count of boundaries within that window. These metrics that both missing and extraneous boundaries are penalized more equitably, providing a robust measure of the model's structural performance.

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[9] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

A.5 Experiment 2 - Query focused summarization

A.5.1 Datasets

We evaluate the performance of our topic-guided query-focused summarisation model across three distinct benchmarks. These datasets are selected to represent varied communicative contexts, ranging from narrative storytelling (ODSum-Story (Rao et al., 2024)) to complex, multi-party meeting transcripts (QMSum and ODSum-Meeting (Rao et al., 2024)).

The ODSum benchmark (Rao et al., 2024) is a recently introduced suite designed for open-domain multi-document summarisation (ODMDS). It consists of two sub-datasets: ODSum-Meeting and ODSum-Story. ODSum-Story is adapted from QMSum, focuses on meeting transcripts, but is structured in an open-domain format where relevant information must be identified and consolidated from an interdependent document index. ODSum-Story is adapted from the SQuALITY dataset (Wang et al., 2022), consisting of public-domain short stories paired with complex, non-factoid queries. Unlike the conversational datasets, ODSum-Story provides narrative inputs, requiring the model to demonstrate a deep understanding of themes and the ability to extract abstractive, query-tailored summaries from dense literary prose. Table **??** provides the comprehensive statistics for each dataset.

A.5.2 Baselines

To evaluate the performance of query-focused summarisation models, we conducted our experiments using various retrieval and answer generation models. For retrieval, we experiment with three different algorithms: BM25 (Robertson & Zaragoza, 2009), Sentence-BERT (SBERT) (Reimers & Gurevych, 2019), and Situated Embeddings (Situated) (Wang et al., 2023). BM25 is a classic, highly effective sparse lexical scoring algorithm widely used in search systems to rank documents by keyword frequency and distribution. For dense semantic retrieval, models like Sentence-BERT (SBERT) encode entire sentences or paragraphs into dense vector representations, enabling retrieval based on meaning rather than just keyword overlap. A more recent advancement is Situated Embeddings (Situated), which are designed to capture the specific context or situation of a query or document to enhance retrieval accuracy.

For answer generation, we evaluated three state-of-the-art open-source LLMs: LLaMA-3.1-8B[10], Gemma-3[11], and Qwen-3-8B, enabling assessment of our method's robustness across diverse model architectures. Following prior work on chunking mechanisms used in RAG (Gao et al., 2024), we compared with three different RAG chunking mechanisms: (1) Fixed-size, splitting documents into contiguous spans of predefined length; (2) Sentence, treating each sentence as an independent retrieval unit; and (3) Semantic, segmenting documents based on embedding similarity between adjacent sentences using an off-the-shelf library.

A.5.3 Metrics

Inspired by (Kirstein et al., 2024), we evaluate the quality of the generated summaries using a complementary suite of metrics that assess lexical overlap, semantic alignment, and referenceless functional utility. By combining these different approaches, we provide a holistic view of model performance that accounts for various types of summary errors, including omissions, hallucinations, and a lack of coherence.

We report the ROUGE scores (Lin, 2004), which remain the standard for measuring the lexical similarity between candidate and reference summaries. ROUGE-1 and ROUGE-2 are used to measure the overlap of unigrams and bigrams, respectively, serving as a proxy for content coverage and informational recall; and ROUGE-L to measure the Longest Common Subsequence (LCS), which captures sentence-level structure and fluency more effectively than individual n-gram matches. While ROUGE is effective for lexical consistency, it often fails to capture semantic paraphrasing. Therefore, we utilise LENS (Maddela et al., 2023). LENS is a supervised metric trained on human judgments that uses a cross-encoder architecture to rank summaries based on their faithfulness and readability. Unlike string-match metrics, LENS can better identify structural disorganisation and subtle semantic deviations that human evaluators typically penalise. To assess the

[10] meta-llama/Llama-3.1-8B-Instruct
[11] google/gemma-3-12b-it

Table 4: Sensitivity of TopicSummRAG to the candidate set size L on QMSum.

L	Precision (↑)	Recall (↑)	F1 (↑)
3	0.386	0.259	0.270
5	0.392	0.267	0.272
8	0.389	0.271	0.271
10	0.387	0.273	0.270

summary's ability to convey the core information of the source document without relying on potentially biased human references, we used BLANC (Vasilyev et al., 2020). BLANC is a referenceless metric that measures the functional quality of a summary by evaluating how much it helps a pre-trained language model perform a masked language modelling (MLM) task on the source document. A high BLANC score indicates that the summary contains sufficient information to help the model "fill in the blanks" of the source text, serving as a strong proxy for document-level information retention.

A.6 Additional Analysis

A.6.1 Retrieval Candidate Size Sensitivity

Table 4 shows that TopicSummRAG exhibits stable performance across a broad range of candidate set sizes. Increasing L slightly improves recall by admitting more candidate segments, while dominance-based adaptive selection prevents the accumulation of irrelevant context for focused queries. As a result, overall F1 remains largely invariant, indicating that retrieval performance is not sensitive to the exact choice of L.

A.6.2 Hyperparameter Sensitivity and Trade-off Analyses

We further analyze the robustness of TopicSummRAG with respect to retrieval granularity and key hyperparameters. These analyses complement the ablation study in Section 4.2 and illustrate how adaptive retrieval balances precision and recall across different operating regimes.

Precision–Recall Trade-off under Adaptive Retrieval Figure 7 (Left) illustrates the precision–recall trade-off for different retrieval strategies corresponding to the ablation settings in Table 3. Fixed retrieval granularities (e.g., k=1 and k=L) exhibit the expected trade-off between precision and recall. In contrast, adaptive-k lies on the Pareto frontier, achieving higher precision than broad retrieval while retaining substantially more recall than single-segment retrieval. This confirms that dominance-based adaptive selection effectively balances coverage and relevance.

Temperature Sensitivity Analysis Figure 7 (Middle) reports performance variation as a function of the softmax temperature τ used for relevance normalization. Performance varies smoothly across the tested range ($\tau \in [0.2, 0.4]$), with τ=0.3 providing a stable operating point. This supports our interpretation of temperature scaling as a mechanism for correcting cosine similarity scale mismatch rather than as a parameter requiring fine-tuning.

Context Window Size Sensitivity Figure 7 (Right) evaluates the effect of the contextual window size W used for cosine-depth boundary detection. Moderate window sizes (W=3–5) yield consistent segmentation performance, while very small or very large windows slightly degrade precision or recall. This trend reflects the trade-off between sensitivity to local transitions and robustness to conversational noise, and indicates that TopicSummRAG is not overly sensitive to window size selection.

Together with the candidate set size analysis in Table 4, these results show that TopicSummRAG is robust to moderate variations in retrieval and segmentation hyperparameters.

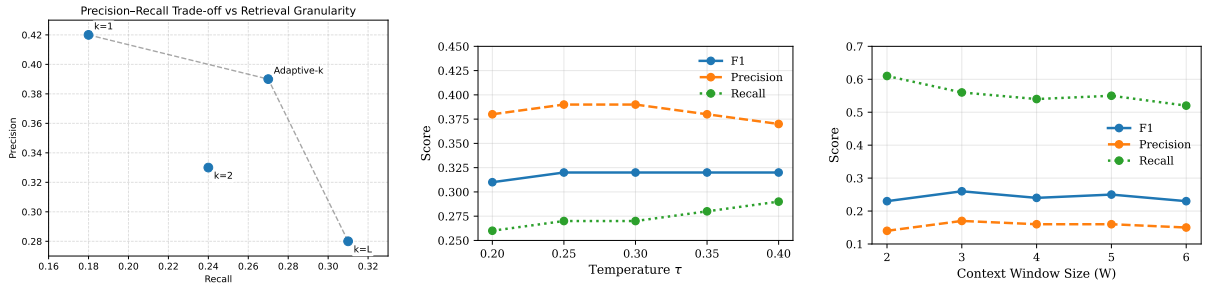

Figure 7: Additional analyses. (Left) Precision–recall trade-off for fixed and adaptive retrieval strategies. (Middle) Sensitivity to softmax temperature τ. (Right) Sensitivity to contextual window size W.

A.7 Query-focused summarization prompt

To generate concise, query-focused summaries from meeting transcripts, we designed a structured prompt (Table 5) that guides the model to extract only relevant content. This prompt ensures that the generated summary remains under 150 words and adheres strictly to the query, producing outputs in a standard JSON format. Such prompting helps maintain consistency and reproducibility in evaluation, and can be directly used for zero-shot summarization with LLMs.

Table 5: Prompt for summary generation

Prompt
You are an expert in query-focused summary from meetings. Given a query and relevant spans from the meeting, your task is to identify query-relevant content from the given span and generate a query-focused summary.
Guidelines
1. Only include content relevant to the query. 2. Keep the query-focused summary under 150 words. 3. Output Format: Return a single JSON object with the key `"query_focused_summary"`. Do not include any additional text or explanations.
Input
Query: Relevant Span:

A.8 G-EVAL summary evaluation

To systematically assess the quality of a query-focused summary, we designed a structured evaluation prompt shown in Table 6. The prompt guides the model to rate a given summary on multiple metrics—relevance, coherence, factuality, and fluency—by comparing it to the story and query. We used *Gemini-2.0-Flash* as the underlying LLM in zero-shot settings.

A.9 ODSum-Story Results

The ODSum-story dataset consists of 4 reference summaries associated for each query, hence, we reported the min-max-mean scores among them in Table 7.

A.10 Error Analysis

As shown in Table 8, errors in query-focused summarisation can arise at multiple stages of the pipeline, including both retrieval and summarisation. Retrieval-level errors occur when the evidence provided to

Table 6: Prompt for G-EVAL

Task Description
You are an expert evaluator for query-focused summarization tasks. Your job is to assess the quality of a query-focused summary against a given story and query using multiple evaluation criteria. The evaluation involves rating the summary on relevance, coherence, factuality, and fluency.
Metric Definition
The definitions of the evaluation criteria are: **Relevance:** How well the summary addresses the information need expressed in the query. **Coherence:** The logical flow, structure, and clarity of the summary. **Factuality:** The correctness and accuracy of the information presented in the summary compared to the story. **Fluency:** The grammaticality, readability, and overall linguistic quality of the summary.
Guidelines
1. Read the query-focused summary carefully and identify its main topic and key points. 2. Compare the summary to the story and check whether it covers the relevant content clearly and logically. 3. Evaluate the summary on each of the four criteria (relevance, coherence, factuality, fluency) on a scale from 1 (lowest) to 5 (highest). 4. Return the results in JSON format with keys "relevance", "coherence", "factuality", and "fluency".
Input
Query: Story: Summary:
Output
{"relevance": X, "coherence": X, "factuality": X, "fluency": X}

the model is incomplete, weakly relevant, or semantically misaligned with the query's information needs. In contrast, summarisation-level errors arise when the model fails to interpret the query intent correctly, inadequately integrates the retrieved evidence, or generates responses that are overly generic, partial, or insufficiently targeted, even when relevant information is available. Crucially, such failures are not necessarily attributable to hallucination; the generated summaries may remain faithful to the retrieved content while still failing to address the query effectively. This distinction motivates the need for a fine-grained error taxonomy that separates retrieval deficiencies from summarisation-specific limitations, enabling more precise diagnosis of query-focused summarisation failures.

A.10.1 Case Study 1

Query What was the two-stage test during prosecutions when discussing the efficacy of the law?

Ground Truth Summary The first stage was to check whether there was sufficient evidence submitted by the police to provide a realistic prospect of conviction. Only if the first stage was satisfied, the second stage came to be considered, which was in the public interest.

The case study Table 9 presents a detailed error analysis for query-focused summarization across six retrieval-summarization configurations: BM25-RAG, BM25-TopicSummRAG, SBERT-RAG, SBERT-TopicSummRAG, Situated-RAG, and Situated-TopicSummRAG. Across these examples, we observe consistent differences in retrieval and summarization quality. BM25-RAG and BM25-TopicSummRAG exhibit notable retrieval shortcomings, with error codes such as RE-04 (Semantic Mismatch) and RE-02 (Partial Retrieval), resulting in summaries that omit or misrepresent the key two-stage prosecution process. Similarly, SBERT-RAG and SBERT-TopicSummRAG retrieved relevant segments more accurately, but summarization issues remain: SU-04 (Overgeneralization) and SU-05 (Over-Inclusion) appear, reflecting summaries that either include excessive background or fail to highlight query-critical details succinctly.

Table 7: Summarization performance on ODSum-Story dataset. *Bold numbers indicate the best performance within each retriever group.*

Retrieval	Chunking	ROUGE-1			ROUGE-2			ROUGE-L		
		Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
				LLaMA-3.1-8B						
BM25	Fixed-size	0.1211	0.1889	0.1537	0.0174	0.0462	0.0307	0.0827	0.1283	0.1046
	Sentence	0.1072	0.1696	0.1375	0.0122	0.0386	0.0243	0.0730	0.1140	0.0925
	Semantic	0.1027	0.1620	0.1318	0.0143	0.0432	0.0280	0.0731	0.1156	0.0931
	TopicSummRAG	0.1978	0.2776	**0.2367**	0.0315	0.0701	**0.0498**	0.1172	0.1667	**0.1408**
SBERT	Fixed-size	0.1215	0.1906	0.1547	0.0158	0.0474	0.0303	0.0834	0.1308	0.1060
	Sentence	0.1233	0.1934	0.1575	0.0180	0.0502	0.0330	0.0861	0.1351	0.1091
	Semantic	0.0930	0.1561	0.1231	0.0138	0.0437	0.0277	0.0685	0.1114	0.0886
	TopicSummRAG	0.2198	0.3014	**0.2594**	0.0346	0.0752	**0.0539**	0.1274	0.1798	**0.1517**
Situated	Fixed-size	0.1352	0.2082	0.1695	0.0225	0.0593	0.0391	0.0913	0.1418	0.1146
	Sentence	0.1929	0.2807	0.2365	0.0347	0.0786	0.0554	0.1223	0.1804	0.1499
	Semantic	0.1140	0.1816	0.1461	0.0198	0.0540	0.0357	0.0809	0.1319	0.1046
	TopicSummRAG	0.2322	0.3172	**0.2735**	0.0389	0.0848	**0.0611**	0.1349	0.1864	**0.1593**
				Qwen3-8B						
BM25	Fixed-size	0.1387	0.2048	0.1708	0.0168	0.0444	0.0298	0.0934	0.1367	0.1139
	Sentence	0.0989	0.1552	0.1263	0.0102	0.0329	0.0208	0.0705	0.1101	0.0894
	Semantic	0.0969	0.1519	0.1239	0.0131	0.0367	0.0241	0.0704	0.1094	0.0889
	TopicSummRAG	0.1913	0.2646	**0.2276**	0.0277	0.0618	**0.0437**	0.1172	0.1631	**0.1395**
SBERT	Fixed-size	0.1495	0.2212	0.1847	0.0191	0.0481	0.0329	0.1000	0.1466	0.1217
	Sentence	0.1428	0.2136	0.1767	0.0174	0.0465	0.0311	0.0963	0.1428	0.1186
	Semantic	0.1026	0.1630	0.1319	0.0147	0.0390	0.0262	0.0741	0.1174	0.0943
	TopicSummRAG	0.2074	0.2829	**0.2438**	0.0296	0.0643	**0.0459**	0.1241	0.1700	**0.1457**
Situated	Fixed-size	0.1624	0.2358	0.1974	0.0225	0.0568	0.0386	0.1042	0.1533	0.1275
	Sentence	0.2010	0.2790	0.2396	0.0288	0.0640	0.0455	0.1220	0.1705	0.1455
	Semantic	0.1290	0.1984	0.1627	0.0185	0.0505	0.0333	0.0882	0.1365	0.1119
	TopicSummRAG	0.2172	0.2972	**0.2571**	0.0324	0.0705	**0.0508**	0.1277	0.1749	**0.1507**
				Gemma3-12B						
BM25	Fixed-size	0.1710	0.2455	0.2077	0.0195	0.0506	0.0339	0.1088	0.1554	0.1309
	Sentence	0.1436	0.2096	0.1758	0.0121	0.0360	0.0232	0.0923	0.1338	0.1120
	Semantic	0.1455	0.2151	0.1795	0.0144	0.0413	0.0274	0.0959	0.1382	0.1164
	TopicSummRAG	0.2216	0.2979	**0.2590**	0.0319	0.0676	**0.0490**	0.1284	0.1721	**0.1493**
SBERT	Fixed-size	0.1747	0.2513	0.2123	0.0197	0.0518	0.0352	0.1094	0.1574	0.1323
	Sentence	0.1602	0.2354	0.1971	0.0190	0.0485	0.0329	0.1050	0.1497	0.1263
	Semantic	0.1340	0.2066	0.1698	0.0149	0.0414	0.0274	0.0908	0.1352	0.1126
	TopicSummRAG	0.2324	0.3107	**0.2716**	0.0337	0.0706	**0.0517**	0.1322	0.1779	**0.1543**
Situated	Fixed-size	0.1810	0.2611	0.2199	0.0245	0.0605	0.0411	0.1139	0.1635	0.1377
	Sentence	0.2258	0.3036	0.2646	0.0345	0.0740	0.0534	0.1312	0.1807	0.1555
	Semantic	0.1561	0.2295	0.1913	0.0204	0.0539	0.0360	0.1016	0.1502	0.1247
	TopicSummRAG	0.2478	0.3303	**0.2876**	0.0383	0.0809	**0.0588**	0.1389	0.1871	**0.1620**

Table 8: Extended Error Taxonomy for Query-Focused Summarization

Code	Stage	Type	Description
Retrieval-Level Errors			
RE-01	Retrieval	Miss	Key evidence required to answer the query is not retrieved.
RE-02	Retrieval	Partial	Retrieved evidence covers only a subset of the information required by the query.
RE-03	Retrieval	Noise	Retrieved segments contain substantial irrelevant or distracting content that obscures useful evidence.
RE-04	Retrieval	Semantic Mismatch	Retrieved content is topically related but lacks the semantic evidence necessary to answer the query.
RE-05	Retrieval	Granularity	Retrieved segments are too coarse or too fine-grained to support effective query-focused summarization.
Summarization-Level Errors			
SU-01	Summarization	Omission	Relevant information present in the retrieved evidence is omitted from the generated summary.
SU-02	Summarization	Query Misinterpretation	The summary reflects an incorrect understanding of the query intent (e.g., answering a different question).
SU-03	Summarization	Evidence-Constrained Failure	The summary remains faithful to the retrieved evidence but fails to answer the query due to insufficient or misaligned information.
SU-04	Summarization	Overgeneralization	The summary provides vague or generic statements instead of specific details required by the query.
SU-05	Summarization	Over-Inclusion	The summary includes excessive background or tangential information, reducing focus on the query.

Situated-RAG demonstrates a stronger retrieval performance by capturing contextually rich segments, though it still shows a combination of retrieval noise (RE-03) and some summarization omissions (SU-02, SU-03). In contrast, Situated-TopicSummRAG consistently delivers the most accurate and query-focused output. Its retrieved chunks encompass the necessary content without extraneous or missing information, allowing the generated summary to fully reflect the two-stage test procedure, including the evidential and public interest stages, and the implications of removing the reasonable chastisement defense. Notably, Situated-TopicSummRAG incurs no reported error codes, highlighting the advantages of topic-chunk retrieval in grounding summarization in relevant and semantically precise evidence. Overall, the analysis indicates that while both BM25 and SBERT retrieval methods can provide partial or noisy inputs, TopicSummRAG-based models, particularly Situated-TopicSummRAG, achieve superior alignment with query requirements, producing summaries that are both comprehensive and faithful to the source evidence.

Table 9: Case Study: Two-stage test

Retrieved Chunk(s)	Generated Summary (Full Text)	Error Code(s)
BM25-RAG Turn 34: *Dawn Bowden AM:* But you would expect that to be something you would.. Turn 38: *Dawn Bowden AM:* That's helpful, thank you. My final question, Chair, is just whether you think there are any other significant cross-border *[truncated]*	The provided transcript does not contain information about a two-stage test during prosecutions related to the efficacy of the law. The discussion revolves around cross-border issues arising from potential legislation, implementation questions, and infrastructure development. Dawn Bowden AM inquired about cross-border issues, while Suzy Davies AM asked specific implementation questions. There's no mention of a two-stage testing process in the provided text.	RE-04; SU-03

Continued on next page

Table 9 – continued from previous page

Retrieved Chunk(s)	Generated Summary (Full Text)	Error Code(s)
BM25-TopicSummRAG *Suzy Davies AM:* Okay, that's really helpful, thank you. *Lynne Neagle AM:* Before we move on, can I just ask—? The Children Act 2004 applies to a child up to the age of 18 .. *Barry Hughes: [truncated]*	The discussion revolved around the applicability of the Children Act 2004, with Lynne Neagle AM initially stating it applies up to age 18, while Barry Hughes countered that it applies up to age 16. This disagreement prompted a clarification of the age range in the context of prosecutions and potential alternatives, as highlighted by Vikki Howells' questions.	RE-02; SU-01
SBERT-RAG Turn 13: *Barry Hughes:* Shall I address that in terms—? Okay. So, you are probably aware, but forgive me if I just explain quickly anyway, when we approach a file of material evidence submitted by the police we apply the code for Crown prosecutors, which has a two-stage *[truncated]*	When approaching a file of material evidence, the Crown Prosecution Service (CPS) applies a two-stage test. The first stage assesses whether there's a realistic prospect of conviction based on sufficient evidence. Only if this first stage is met does the CPS proceed to consider whether prosecution is in the public interest. The police apply the same code before referring cases to the CPS, although disagreements sometimes occur. Even in cases involving seemingly minor incidents like a 'light smack', the evidential stage must be satisfied before considering public interest. If the defence of reasonable chastisement is removed, cases may be referred to the CPS, though police might still opt not to prosecute if it's not in the public interest.	RE -03; SU-04
SBERT-TopicSummRAG *Vikki Howells AM:* Thank you, Chair. You've already touched quite a bit on the issue of the public interest test *[truncated]*	The discussion centered on the two-stage test during prosecutions and how it would be affected by the proposed legislation. Currently, the CPS considers a two-stage process: first, assessing evidential stage, and then the public interest test. The removal of the reasonable chastisement defence would mean the public interest test is applied sooner, particularly in cases of repeated behavior. The CPS would assess proportionality, considering factors like the age of the offender and victim, circumstances, and impact. They do not anticipate a significant increase in prosecutions due to the law, as they only prosecute when there's a realistic prospect of conviction and it's in the public interest. While there's a shortage of registered intermediaries, it's not expected to be a significant barrier.	SU-05

Continued on next page

Table 9 – continued from previous page

Retrieved Chunk(s)	Generated Summary (Full Text)	Error Code(s)
Situated-RAG Turn 13: *Barry Hughes:* Shall I address that in terms—? Okay. So, you are probably aware, but forgive me if I just explain quickly anyway, when we approach a file of material evidence submitted by the police we apply the code for Crown prosecutors, which has a two-stage *[truncated]*	The Crown Prosecution Service (CPS) applies a two-stage test when assessing cases involving physical discipline of children. The first stage assesses whether there's sufficient evidence to realistically prospect a conviction. Only if this initial stage is met does the CPS proceed to the second stage, considering public interest factors. Currently, the defense of reasonable chastisement often prevents cases from reaching the evidential stage. If the proposed Bill removes this defense, it is anticipated that police will refer more cases to the CPS, though not all referrals will lead to prosecution. The CPS will independently apply the same two-stage test, considering factors like the severity of the incident, the child's age, and any history of such incidents. The public interest test will consider proportionality, taking into account various factors before a charging decision is made.	SU-03; RE-03; SU-02
Situated-TopicSummRAG Chunk 1: *Vikki Howells AM:* Thank you, Chair. You've already touched quite a bit on the issue of the public interest test *[truncated]* Chunk 2: *Lynne Neagle AM:* Good afternoon, everyone. Welcome to the Children, Young People and Education Committee *[truncated]*	The two-stage test during prosecutions, as outlined in the Code for Crown Prosecutors, will remain in place. The first stage assesses whether there's a realistic prospect of conviction, while the second stage considers whether prosecution is in the public interest. The proposed legislation in Wales, removing the defense of reasonable punishment, could lead to more frequent referrals to the CPS, particularly for cases.	None

A.10.2 Case Study 2

Query. Summarize the discussion about the divergence between the law in England and Wales.

Ground Truth Summary. Barry Hughes indicated that the Bill would impact England and their people, although it needed a period. Moreover, Barry thought there was no cross-border issue from the potential legislation, and the draft changing to CPS guidance would not be helpful for scrutinisers to decide on the law in Wales.

Case Study 2 (Table 10) highlights differences in query-focused summarization across six configurations regarding the divergence of criminal law in England and Wales. BM25-RAG shows retrieval misses (RE-01) and evidence-constrained summaries (SU-03), while BM25-TopicSummRAG retrieves partial evidence (RE-02) leading to omitted details (SU-01). SBERT-RAG and Situated-RAG capture relevant content more accurately but still overgeneralize (SU-04) or misinterpret query-critical nuances (SU-02). SBERT-TopicSummRAG improves query alignment but omits some procedural detail. Situated-TopicSummRAG achieves the most accurate and comprehensive summary, retrieving all key content with only minor overgeneralization (SU-04). Overall, TopicSummRAG-based retrieval demonstrates superior ability to produce summaries faithful to source evidence and fully aligned with query requirements.

Table 10: Case Study: Divergence between the law in England and Wales

Retrieved Chunk(s)	Generated Summary (Full Text)	Error Code(s)
BM25-RAG Turn 38: TopicSummRAG Dawn Bowden AM: That's helpful, thank you. My final question, Chair, is just whether you think there are any other significant cross-border issues that arise from this potential legislation. </content>	The discussion touched upon potential cross-border issues arising from the legislation, particularly concerning divergence between the law in England and Wales. Specific questions around implementation were raised, and the suitability of the legislation being 'fit for purpose' was briefly addressed. No specific details about the nature of the divergence were provided in the span.	RE-01; SU-03
BM25-TopicSummRAG *Lynne Neagle AM:* Siân. Sian Gwenllian AM: My point is about the criminal law in Wales and England diverging in this particular instance, and we've been told that this is probably the first time it's happened.	The discussion highlighted a divergence between criminal law in England and Wales, particularly regarding offenses like putting electronic collars on pets and picking cockles in West Wales, which are offenses in Wales but not in England. A key point was the potential for children to receive different levels of protection based on their age and location. While the CPS in England will need to be aware of these changes, the extent of awareness required will be less than in Wales due to the regular interchange of prosecutors.	RE-02; SU-01
SBERT-RAG Turn 17: *Lynne Neagle AM:* We've got some questions now on the divergence between the law in England and Wales from Dawn Bowden. Turn 19: *Barry Hughes:* Can I give you two examples? One of the implications is this—and it's fairly ...	The discussion centered on the divergence between English and Welsh law, particularly concerning the removal of the reasonable punishment defense for assaults on children. Examples illustrated the divergence, such as a family traveling from England to Wales where minor smacks permissible in England could be illegal in Wales. The CPS intends to issue policy guidance and charging standards to address this divergence.	RE-02; SU-04
SBERT-TopicSummRAG *Lynne Neagle AM:* Siân. *Sian Gwenllian AM:* My point is about the criminal law in Wales and England diverging in this particular instance, and we've been told that this is probably the first time it's happened.	The discussion centered on the divergence between criminal law in England and Wales, particularly regarding the potential removal of a defense for assaults on children. CPS staff awareness and guidance updates were highlighted. Cases in Wales would likely pass the evidential stage more often than in England due to absence of the defense.	SU-02

Continued on next page

Table 10 – continued from previous page

Retrieved Chunk(s)	Generated Summary (Full Text)	Error Code(s)
Situated-RAG Turn 17: TopicSummRAG *Lynne Neagle AM:* We've got some questions now on the divergence between the law in England and Wales from Dawn Bowden. Turn 19: *Barry Hughes:* Can I give you two examples? One of the implications is this—and it's fairly ...	The discussion centered on the divergence between English and Welsh law, particularly regarding the removal of the defense for assaults upon children. Awareness among prosecutors, policy guidance, and potential impact on individuals' life prospects were key discussion points.	RE-02; SU-04
Situated-TopicSummRAG Chunk 1: *Barry Hughes:* If this Bill goes through, then it will remove a defence, which will make it likely that the police will give this more consideration for referral to the CPS than beforehand, ... Chunk 2: *Dawn Bowden AM:* Okay. Now, we've just talked about the divergence of criminal law, and this is probably the first ...	The discussion centered on the divergence between English and Welsh law, particularly concerning the potential removal of a defense for assaults on children. Barry Hughes highlighted that the removal of this defense in Wales would likely lead to more cases reaching the CPS compared to England. The CPS plans to update guidance and raise staff awareness, with phased implementation over potentially two years.	SU-04

A.11 Significant t-test

A.11.1 Per-Sample ROUGE-L Analysis and Statistical Significance

To quantify the improvement of Situated TRAG over baseline, we conducted a paired t-test on ROUGE-L scores computed for each document in our evaluation set.

The paired t-test results are as follows:

- t-statistic: 14.8245
- p-value: 5.38×10^{-36}
- Mean Δ ROUGE-L (Situated TRAG – baseline): 0.0652
- 95% Confidence Interval: [0.0565, 0.0738]

The extremely low p-value indicates that the improvement in ROUGE-L is highly significant, confirming that Situated TRAG consistently outperforms the baseline in capturing query-focused summary quality.

Per-Sample Analysis Figure 8 shows the distribution of per-document improvements (Δ ROUGE-L = Situated TRAG - baseline). The majority of documents exhibit positive gains, indicating consistent improvements across the dataset.

Overall, both the statistical testing and per-sample analysis demonstrate that Situated TRAG provides significant and consistent improvements in query-focused meeting summarization compared to a baseline.

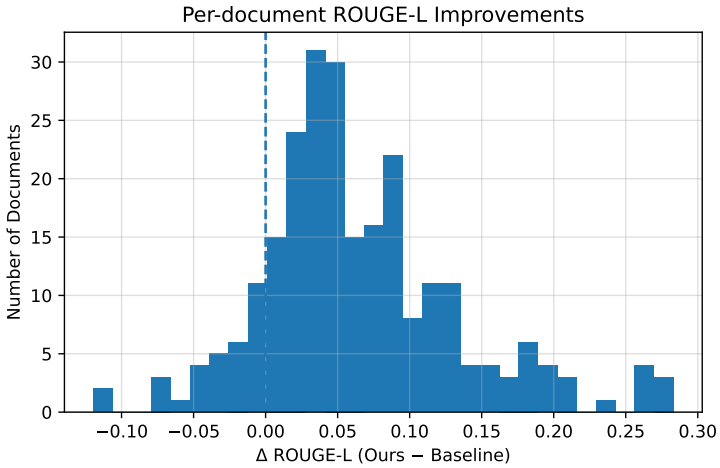

Figure 8: Distribution of per-document ROUGE-L improvements (Δ ROUGE-L = Situated TRAG − baseline). Positive values indicate documents for which Situated TRAG outperforms the baseline. The distribution is skewed toward positive gains, suggesting that Situated TRAG consistently improves ROUGE-L scores across most documents.