Topic-Guided Reinforcement Learning with LLMs for Enhancing Multi-Document Summarization

Anonymous ACL submission

Abstract

A key challenge in Multi-Document Summarization (MDS) is effectively integrating information from multiple sources while maintaining coherence and topical relevance. While Large Language Models (LLMs) have shown impressive results in single-document summarization, their performance on MDS still leaves room for improvement. In this paper, we propose a topic-guided reinforcement learning approach to improve content selection in MDS. We first show that explicitly prompting models with topic labels enhances the informativeness of the generated summaries. Building on this insight, we propose a novel topic reward within the Group Relative Policy Optimization (GRPO) framework to measure topic alignment between the generated summary and source documents. Experimental results on the Multi-News and Multi-XScience datasets demonstrate that our method consistently outperforms strong baselines, highlighting the effectiveness of leveraging topical cues in MDS.

1 Introduction

Multi-Document Summarization (MDS) aims to generate a concise and coherent summary that captures the salient information from a collection of related documents. While recent advances in Large Language Models (LLMs) and prompting strategies have significantly improved the performance of abstractive summarization systems, existing MDS methods still struggle to maintain content relevance, coherence (Belem et al., 2024), and topic consistency (Amar et al., 2023), especially when synthesizing information across multiple sources (Liu et al., 2024; Lior et al., 2024).

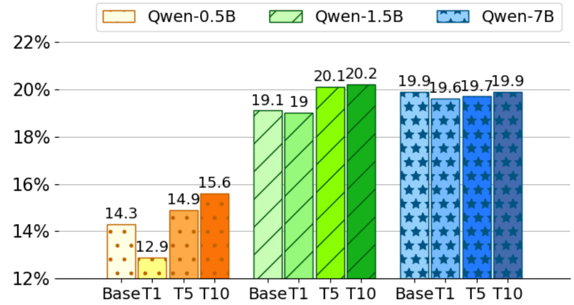
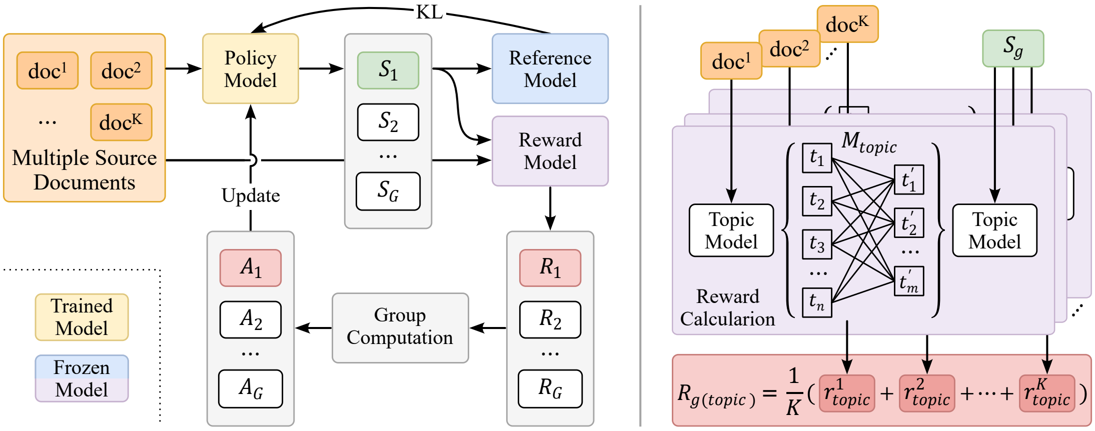


Figure 1: Performance on Multi-News (Fabbri et al., 2019) using prompting (Base) and topic-incorporated prompting (Tn; n means number of topic labels) with Qwen2.5-series model (Qwen et al., 2025). The geometric mean of Rouge-1/2/L scores are reported. Topic key words are previously generated using a *teacher* model: Qwen2.5-7B. We see that topic-enhanced instruction (T5 and T10) improves small LLMs' (0.5B and 1.5B) performances over standard prompt (Base).

One important yet relatively underexplored direction in MDS is the incorporation of high-level discourse information to guide the summarization process. Topics offer a global discourse structure that can help models identify salient content, resolve ambiguity, and enhance coherence in the generated summaries (Haghighi and Vanderwende, 2009; Ouyang et al., 2007). Early work incorporated topic distributions as auxiliary features to enrich word and sentence representations, either via topic models or graph-based approaches (Wei, 2012; Narayan et al., 2018; Wang et al., 2020). However, these methods typically operate at the token or sentence level and do not fully leverage topic signals as explicit guidance. More recent efforts have attempted to better align topic modeling with the summarization objective—for example, by using latent topics to pre-select salient sentences in extractive settings (Cui et al., 2020), or by jointly learning topic representations and summarization (Cui and Hu, 2021). Related work has also explored creating intermediate plans to guide summarization, such as creating *Entity Chains* as key phrases (Narayan et al., 2021) or building question-answering blueprints (Narayan et al., 2023). Despite these advances, the explicit use of topic labels as prompts or rewards to guide multi-document summarization remains largely unexplored.



(a) GRPO Training

(b) Topic-Guided Reward

Figure 2: Multi-Document Summarization training (a) using our proposed Topic-Guided reward (b), with GRPO (Shao et al., 2024). Every input data contains K source documents (doc^1, . . . , doc^K). For each document doc^k, we use a topic model to extract n number of key topic phrases $\{t_1, t_2, \ldots, t_n\}$. Similarly, we extract m topic phrases $\{t'_1, t'_2, \ldots, t'_m\}$ from each generated summary S_g. We construct a topic similarity matrix M^k_g by comparing the n and m topic phrases from each source document-summary pair, from which we compute a topic alignment score $r^k_{g\,(\text{topic})}$. We average the alignment scores over all K document-summary pairs to derive the overall topic-guided reward $R_{g\,(\text{topic})}$, which is then used to calculate the group advantage A_g for updating the policy model.

In this work, we investigate the role of explicit topic guidance in enhancing generic MDS. Different from previous studies that incorporate topic distributions or learns latent topics via neural topic modeling, we propose a more direct and interpretable strategy: guiding summarization models using topic phrases explicitly extracted from the source documents. We begin with a simple yet insightful observation: prompting LLMs with extra topic information improves MDS quality in terms of informativeness. Figure 1 shows that when small LLMs (Qwen2.5-0.5B and 1.5B) are applied to summarization tasks, they show notable improvements if prompted with topic labels ("T5" and "T10"), compared to using standard summarization prompt ("Base"). This motivates us to go beyond static prompting and incorporate topic awareness more directly into the training objective.

To this end, we introduce a novel reference-free topic-reward function that quantifies how well a generated summary aligns with its intended topics derived from each source document, see Figure 2 for an overview. Our key assumption is that increasing the topical similarity between the generated summary and source documents will in turn improve the quality of summary generations. Accordingly, our reward is defined with respect to the improvements from (1) coverage: how well the generated summary covers important topics in source documents, and (2) precision: how relevant the topics in summary are to the source documents. The final reward signal is a harmonic mean, which is then integrated into the Group Relative Policy Optimization framework (Shao et al., 2024) to enable reinforcement learning with topic-guided feedback.

Specifically, we employ Qwen2.5-7B (Qwen et al., 2025) within the reward model to generate topic labels for a given document–either a source article or a summary–while using the smaller Qwen2.5-0.5B model as the policy model. This setup also mirrors a knowledge distillation paradigm where the larger language model transfers topic-related knowledge to the smaller model to guide its learning process. We evaluate our method on two widely-used datasets: Multi-News (Fabbri et al., 2019) and Multi-XScience (Lu et al., 2020), and demonstrate that our topic-aware training strategy leads to consistent improvements over standard and Reinforcement Learning from Human Feedback (RLHF)-guided baselines, as measured by both informativeness metrics (e.g., ROUGE, LLM score) and topic alignment evaluation.

In summary, (1) we show that using topic information improves MDS performance, both via prompting and LLM-based RL; (2) we introduce a novel topic reward to measure source-summary discourse alignment, which is integrated into GRPO to perform topic-guided summarization; (3) empirical results indicate that topic-level signals represent a valuable yet underexploited form of supervision, yielding even stronger performance when combined with reference-based rewards like ROUGE.

2 Related Work

Multi-Document Summarization (MDS) Early approaches for MDS relies on extractive methods that rank and select salient sentences across documents (Nenkova and Vanderwende, 2005; Erkan and Radev, 2004). More recent work has shifted toward neural abstractive models that can generate coherent and fluent summaries from scratch (Liu and Lapata, 2019; Zhang et al., 2020; Ma et al., 2022). However, these models often face challenges in maintaining factual accuracy and topical consistency due to the complexity of aggregating information from multiple sources. Several techniques have been proposed to address this, such as hierarchical encoding (Liu et al., 2018), guided decoding strategies (Pasunuru et al., 2021), and multi-granularity control over an extract-then-summarize pipeline (Zhang et al., 2024). Despite these advances, the integration of high-level discourse information such as topics remains underexplored.

Discourse-Guided Summarization Although topic modeling has been widely used for document-level content understanding, its application to summarization has been relatively limited (Cui and Hu, 2021). Haghighi and Vanderwende (2009) used LDA-style probabilistic topic models (Blei et al., 2003) to select topic-relevant sentences and showed improvement in terms of redundancy; Cohan et al. (2018) and Wang et al. (2020) introduced discourse-level and topic-aware attention mechanisms to enhance long document summarization.

Another line of work involves discourse-level planning, where models generate summaries conditioned on given keywords (He et al., 2022; Dou et al., 2021), entities (Narayan et al., 2021), or high-level concept (Zhong et al., 2021). These approaches aim to control the focus of the summary based on user intent or query, while our work focuses on generic summaries that holistically represent the source content using topical information.

Reinforcement learning (RL) for Summarization RL methods has been applied to summarization more broadly to optimize non-differentiable objectives, such as ROUGE (Ranzato et al., 2016; Paulus et al., 2018; Narayan et al., 2018) or human preferences (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022), addressing the inherent mismatch between training objectives and evaluation criteria. Other approaches incorporate task-specific rewards, e.g., Wu and Hu (2018) learned models of coherence from existing text and used them as RL rewards for summarization; Gao et al. (2019) built an interactive summarization tool by applying reward learning to one article at a time; Pasunuru and Bansal (2018) incorporated entailment-based consistency rewards to improve the saliency of a good summary. Our work builds on this line by introducing a novel topic-level reward that explicitly encourages topical alignment between generated summaries and source documents–a dimension not adequately addressed by existing metrics.

3 Incorporating Topic Labels into MDS via Prompting

Topic phrases succinctly capture essential information from source documents, providing effective high-level guidance for summarization–our motivation aligns with prior work on *Entity Chains* (Narayan et al., 2021), which utilized ordered sequences of entities as intermediate representations to plan and ground abstractive summary generation. However, unlike the controlled entity sets used in entity chains, we treat our topics as open-ended keywords and phrases. Additionally, rather than incorporating entity generation directly into conditional summarization, we adopt a two-step framework where the topic extraction model is separate from the summarization model. This modular design enables independent analysis of topic extraction's impact on summarization quality (see §6.4), and provides the flexibility to incorporate more advanced topic models in future experiments.

In this section, we conduct experiments in a zero-shot setting, carefully designing prompts and configurations to best leverage topic-augmented MDS.

Prompting with Topics Formally, given a set of source documents with corresponding topic labels, we prompt a LLM for summarization as follows:

$$P(S|doc^1, T_{doc^1}, \ldots, doc^K, T_{doc^K}; \theta), \quad (1)$$

where T_{doc^k} denotes topic labels for document k, and θ represents the LLM parameters. We append each set of topic labels immediately after its corresponding document, providing explicit topical guidance to assist the summarization model. This resembles the *summary-level entity plans* introduced in Narayan et al. (2021), but extends naturally to multiple document-topic pairs, a format we found consistently more effective than an aggregated-topic version in pilot experiments. Detailed prompt examples are provided in Appendix A.

***Teacher-Supervision* Mode** We examine LLM capabilities by comparing their performance of varying scales (Qwen2.5-0.5B, 1.5B, and 7B). Unsurprisingly, the largest summarization model (7B) achieves the highest baseline performance (average ROUGE 19.9), significantly surpassing smaller models (see "Base" in Figure 1). We employ a teacher-supervision mode, where the larger 7B model explicitly provides topic guidance for the smaller models (0.5B and 1.5B). Under this setting, smaller LLMs clearly benefit from improved topical information provided by the teacher model. However, the 7B model itself, which inherently possesses strong topical modeling capabilities, experiences no gains from self-generated topic labels.

Number of Topic Labels We also explore how the number of topic labels impacts summarization effectiveness, comparing summaries guided by 1, 5, or 10 topics. A single topic label overly constrains summarization, leading to poorer performance across all models ("T1" in Figure 1). On the other hand, summarization quality notably improves when using more labels ("T5" and "T10"), particularly for smaller models.

These findings together suggest that employing richer topic signals through teacher-supervised extraction is beneficial, motivating us to incorporate topic information into the learning process.

4 LLM Reinforcement Learning for MDS

Based on the above observations, we propose a novel topic-guided reward (§4.1) designed to maximize semantic similarity between generated summaries and source documents, coupled with a length penalty (§4.2) to better control the generation length. We implement these rewards using an inverse standard deviation weighting strategy (§4.3) through the recent Group Relative Policy Optimization (GRPO) framework (§4.4). See Figure 2 for the overview of our pipeline.

4.1 Topic-Guided Reward

A key contribution of our approach is a **Topic-F1 reward** metric that can effectively capture the semantic alignment between summaries and their respective source documents. We utilize a two-step embedding and matching procedure to quantify coverage and precision of topics.

For one data input $d = \{doc^1, doc^2, \ldots, doc^K\}$, we first apply the Qwen2.5-7B model to extract a set of topic labels $T_{doc} = \{t_1, t_2, \ldots, t_n\}$ from each source document doc^k. Specifically, we set topic number $n = |T_{doc}| = 10$ for Multi-News (Fabbri et al., 2019) and $n = |T_{doc}| = 5$ topics for Multi-XScience (Lu et al., 2020). These values are defined to align with the average number of sentences per summary in the training data (Table 1). Each topic label–may be a single word or a short phrase–is converted into a dense embedding using the SentenceTransformer model `all-mpnet-base-v2` (Reimers and Gurevych, 2019). We select this model due to its compact size and proven effectiveness in generating high-quality sentence embeddings, adding minimal computational overhead in training.

Given a generated summary S_g, we similarly extract and embed its topics $T_{\text{sum}} = \{t_1, t_2, \ldots, t_m\}$. We construct a similarity matrix M, whose entries M_{ij} represent the cosine similarity between topic embeddings of each pair of topic phrases from source document and generated summary:

$$M_{ij} = \frac{\mathbf{e}_{\text{doc,i}} \cdot \mathbf{e}_{\text{sum,j}}}{|\mathbf{e}_{\text{doc,i}}||\mathbf{e}_{\text{sum,j}}|}, \tag{2}$$

where $\mathbf{e}_{\text{doc,i}}$ and $\mathbf{e}_{\text{sum,j}}$ represent embeddings for the i^{th} document topic and j^{th} summary topic, respectively. Note that the number of extracted topics from the source document and the generated summary may differ, as summaries are typically much shorter than source documents. We set the number of topics $m = |T_{\text{sum}}| = 5$ for both datasets.

Then, we define **Coverage** as the average of the maximum similarity scores between each source topic and its most similar summary topic. Conversely, **Precision** is defined as the average of the maximum similarity scores between each summary topic and its most similar source topic:

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} \max_{j=1,2,\ldots,m} (M_{ij}), \tag{3}$$

$$\text{Precision} = \frac{1}{m} \sum_{j=1}^{m} \max_{i=1,2,\ldots,n} (M_{ij}). \tag{4}$$

Finally, we calculate the **harmonic mean** of coverage and precision to derive our topic-guided reward r_{topic}. This metric is computed pairwise for every source document-summary pair, encouraging generation of generic summaries that consistently capture key semantic elements across multiple documents. The final reward score R_{topic} is obtained by averaging the r_{topic} values across all document-summary pairs of one data point. Preliminary experiments revealed that computing topic rewards on a pairwise basis consistently outperformed approaches that first merged topics across all documents before comparison, motivating our choice of topic alignment calculation.

4.2 Length-Penalty Reward

As recent research shows, LLMs often fail to respect desired length constraints specified in prompts (Stiennon et al., 2020; Wang et al., 2024). To mitigate excessive long (or short) output, we introduce a **token-level length reward** designed to penalize deviations from the target length. To determine the number of tokens, we use the tokenizer associated with the reference model–specifically, the Qwen2.5-0.5B model in our case.

Formally, the length reward R_{len} is defined as:

$$R_{\text{len}} = \exp\left(-\frac{|L_{\text{exp}} - L_{\text{sum}}|}{L_{\text{exp}}}\right), \tag{5}$$

where L_{exp} represents the desired summary length and L_{sum} the generated summary length. We compute L_{exp} on a small validation set, with its size tunable to reflect user preferences.

In our pilot experiments, we evaluated both sentence-level and token-level approaches for length penalty. The results showed that token-level control led to significantly better adherence to the target length, effectively preventing summaries from becoming excessively long (up to five times the target length observed in initial trials).

4.3 Reward Weighting

Our reward formulation can be viewed within the broader Multi-Objective Reinforcement Learning (MORL) framework, where multiple objectives–topic precision, coverage, and length constraints–must be simultaneously balanced. Inspired by the MORL literature (Roijers et al., 2013; Van Seijen et al., 2017) and adaptive weighting strategies such as leveraging reward variance (Kendall et al., 2018), we adopt an **inverse standard deviation weighting** scheme to stabilize training signals. Given reward signals R_r with standard deviations σ_r which we obtain from a mini-batch (approx. 5% of training set), where r refers to the reward type, the initial weights are defined as:

$$w_r = 1/\sigma_r \tag{6}$$

Additionally, following common practice (Stiennon et al., 2020), we apply an emphasis factor of 2 to the topic-guided reward to reflect domain-specific priorities. This factor is a tunable hyperparameter, selected based on development set performance. The final weights are normalized across all reward types:

$$w_r^{\text{norm}} = \frac{w_r \times \text{factor}_r}{\sum_k (w_k \times \text{factor}_k)}, \tag{7}$$

where $\text{factor}_{\text{topic}} = 2$ and $\text{factor}_{\text{len}} = 1$. In further experiments, we incorporate the reference-based ROUGE reward alongside our reference-free topic-F1 reward, assigning equal weighting to both. This strategy efficiently balances multiple reward components and dynamically emphasizes key metrics.

4.4 GRPO Training

To integrate our weighted reward into GRPO training (Shao et al., 2024), we construct a scalar value R_{total} which combines topic-F1 and length rewards:

$$R_{\text{total}}(S_g) = \sum_r w_r^{\text{norm}} R_r(S_g). \tag{8}$$

The GRPO algorithm computes relative advantages of R_{total} within a group of G sampled completions, i.e., generated summaries:

$$A_{\text{g}}^{\text{GRPO}} = \frac{R_{\text{total}}(S_g) - \frac{1}{G}\sum_{g=1}^{G} R_{\text{total}}(S_g)}{\text{std}_{g=1,2,\ldots,G}(R_{\text{total}}(S_g))}. \tag{9}$$

Given this advantage estimation, the training objective is to optimize the policy (π) parameters θ by maximizing a clipped surrogate objective:

$$\begin{aligned} L^{\text{GRPO}}(\theta) = \mathbb{E}_{S_g \sim \pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G} \sum_{g=1}^{G} \min\Big(r_g(\theta) A_g, \\ \text{clip}(r_g(\theta), 1-\epsilon, 1+\epsilon) A_g \Big) \Bigg] \\ - \beta \cdot \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \end{aligned} \tag{10}$$

with probability ratio $r_g(\theta) = \frac{\pi_\theta(S_{\text{g}}|d)}{\pi_{\theta_{\text{old}}}(S_{\text{g}}|d)}$ and a KL penalty to regularize policy updates. Note however that our reward design is agnostic to specific RL algorithms, we adopt the GRPO framework due to its recent success (Shao et al., 2024; Guo et al., 2025) and computational efficiency by removing the value model.

5 Experimental Setup

5.1 Datasets

We choose two popular MDS datasets whose source documents and summaries different along multiple facets such as length and abstractiveness. The key statistics of datasets are shown in Table 1. **(1) Multi-News** (Fabbri et al., 2019) is one of the most widely used MDS datasets in news domain. It

	Multi-News	Multi-XSci
Nb. train data	$44,972$	$30,369$
Nb. test data	$5,622$	$5,093$
Nb. refs per summ	2.8	4.4
Avg. words / sents in docs	$2,103$ / 28	942 / 33
Avg. words / sents in summ	263 / 10	116 / 5
% novel unigrams	17.8	57.1
% novel bigrams	42.3	81.8

Table 1: Key statistics of Multi-News (Fabbri et al., 2019) and Multi-XScience (Lu et al., 2020). These numbers show variance in the size of source documents (references per summary, avg. words and sentences) and difference in gold summary properties (novel n-grams).

contains in average 2.7 source documents per summary with relatively long documents. **(2) Multi-XScience** (Lu et al., 2020) comprises the abstract of a query paper and those of its cited papers as input, with the goal of generating a related work paragraph. On average, it includes 4.4 source documents and has highly abstractive summaries, making it particularly challenging for MDS models.

5.2 Evaluation Metrics

We report several complementary metrics that examine different aspects of the generated summaries. To assess summary *informativeness*, we use lexical overlap metrics (e.g., ROUGE; (Lin, 2004)), along with embedding-based semantic similarity measures including BERTSCORE (short in BERT; Zhang et al., 2019) and LLM2VEC SCORE (short in LLM2V; BehnamGhader et al., 2024). The LLM2V metric we use is built upon the `Meta-LLaMA-3-8B` model fine-tuned (Grattafiori et al., 2024) with unsupervised contrastive learning (Gao et al., 2021). Additionally, we examine *topical alignment* via COVRATIO and PRERATIO, reflecting respectively the coverage and precision of extracted topics between the summary and source documents.

5.3 Model Comparisons

We primarily use Qwen series for our experiments (Qwen et al., 2025). For all model variants, Qwen-2.5 0.5B-Instruct is used as the policy model in RL training. For reward calculation, we compare different sizes and types of reward model. For all RL-training, we include the length penalty described in §4.2. We compare the following variants:

(1) RL-Trained, Topic-reward: Our proposed method, training a policy model (0.5B) with topic-F1 reward and GRPO. We include $\text{RL}_{\text{TOPIC-7B}}$ which leverages Qwen-2.5 7B-Instruct model as topic extractor, and explore a smaller variant $\text{RL}_{\text{TOPIC-0.5B}}$ with 0.5B model for topic extraction.

(2) RL-Trained, Human-feedback: We compare against a reward model trained to predict human preference from OpenAssistant[1] (`deberta-v3-large-v2`): $\text{RL}_{\text{HUMAN-FEEDBACK}}$.

(3) Base: We also compare against Qwen2.5 0.5B model evaluated in a zero-shot setting, both with topic labels provided by Qwen 7B in the prompt ($\text{BASE}_{\text{TOPIC-7B}}$) and without any topic information (BASE). $\text{BASE}_{\text{TOPIC-7B}}$ approximates the *Entity Chains* (Narayan et al., 2021) within LLM.

(4) Supervised Fine-Tuning (SFT): We fine-tune Qwen2.5-0.5B-Instruct model for summary generation with the SFT objective.

(5) RL-Trained, Reference-based: Finally, we implement ROUGE-reward using the mean of ROUGE-1/2/L within GRPO: RL_{ROUGE}, and benchmark with our model which uses a combination of topic and ROUGE rewards: $\text{RL}_{\text{TOPIC-7B+ROUGE}}$.

5.4 Implementation Details

We adapt the TRL library (von Werra et al., 2020) for GRPO training. Most of our experiments are conducted using 8 x NVIDIA A100 40GB GPUs, where one GPU is dedicated for rollout, one for topic generation, and the rest for GRPO training. We set the number of generations to 8, per-device train batch size to 4, gradient accumulation steps to 21, and KL coefficient to 0.04. For rollout and topic generation, we using vLLM[2] for accelerated inference, with more details in the Appendix B.

6 Results

6.1 Main Results

In Table 2, we report results comparing our methods with baselines. Across both datasets, our method consistently outperforms all baselines in terms of summary informativeness. Specifically, on Multi-News, $\text{RL}_{\text{TOPIC-7B}}$ achieves superior embedding-based similarity scores (.845 for BERTSCORE and .798 for LLM score) compared to $\text{RL}_{\text{HUMAN-FEEDBACK}}$ (.819 BERTSCORE and .706 LLM score). Even our smaller topic-guided variant, $\text{RL}_{\text{TOPIC-0.5B}}$, notably surpasses this baseline in all metrics, illustrating the robustness and scalability of our topic-guided reward framework. A similar trend is observed for Multi-XScience.

[1] https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2.

[2] https://github.com/vllm-project/vllm

	Model	RM*	Overlap-Based				Similarity-Based		Topic Alignment	
			Rouge-1	Rouge-2	Rouge-L	Rouge-M	BERT	LLM2V	COVRATIO	PRERATIO
News	BASE	-	27.22	7.28	15.03	14.31	.842	.721	.513	.622
	BASE$_{\text{TOPIC-7B}}$	7B	28.62	8.60	15.83	15.73	.844	.733	.521	.632
	RL$_{\text{HUMAN-FEEDBACK}}$	0.3B	33.07	6.99	17.29	15.58	.819	.706	.492	.583
	RL$_{\text{TOPIC-0.5B}}$ (ours)	0.5B	38.63	10.72	18.81	19.82	**.845**	.793	.536	.672
	RL$_{\text{TOPIC-7B}}$ (ours)	7B	**39.62**	**10.97**	**18.97**	**20.20**	**.845**	**.798**	**.540**	**.676**
XScience	BASE	-	25.05	4.16	13.47	11.19	.822	.637	.490	.480
	BASE$_{\text{TOPIC-7B}}$	7B	25.62	4.09	13.93	11.34	.828	.655	.482	.479
	RL$_{\text{HUMAN-FEEDBACK}}$	0.3B	26.78	2.90	13.87	10.25	.832	.622	.506	.507
	RL$_{\text{TOPIC-0.5B}}$ (ours)	0.5B	29.47	4.79	15.90	13.09	.835	.721	.548	.549
	RL$_{\text{TOPIC-7B}}$ (ours)	7B	**30.45**	**5.38**	**16.26**	**13.86**	**.847**	**.741**	**.554**	**.560**

Table 2: Model performance on two MDS datasets. We report ROUGE scores (Rouge-1/2/L/Mean) (Lin, 2004), BERTSCORE (Zhang et al., 2019), and LLM2V score (BehnamGhader et al., 2024), computed against gold summary. We assess topic alignment using coverage ratio (COVRATIO) and precision (PRERATIO). Best score per column is in **bold** and second best underlined. All models used for summary generation are of size 0.5B. RM* shows the size of reward model in RL settings and topic extraction model in zero-shot setting.

In addition to informativeness metrics, we introduce a novel topic alignment assessment that directly evaluates the semantic alignment between generated summaries and source documents, independently of gold reference summaries. Our topic-guided models demonstrate significant improvements on both datasets, achieving increases of 2-7 points in Coverage and 4-8 points in Precision compared to baseline models. These consistent enhancements highlight the value of integrating direct topical guidance into the summarization process.

6.2 Results with SFT and Rouge-Based RL

We further evaluate our combined reward strategy (RL$_{\text{TOPIC-7B+ROUGE}}$) against reference-based approaches, specifically supervised fine-tuning (SFT) and RL with ROUGE rewards. As shown in Table 3, our model consistently surpasses these baselines in both similarity metrics and topic alignment scores.

On Multi-News, while SFT achieves slightly higher ROUGE scores due to its direct token-prediction training, our method notably excels in capturing semantic similarity. In the more challenging Multi-XScience dataset–characterized by a larger number of source documents and highly abstractive summaries, as shown in Table 1–our RL model demonstrates clear superiority across all evaluated metrics. This highlights RL's capacity to develop comprehensive summarization strategies beyond simple token imitation. Interestingly, the RL approach using solely ROUGE as reward (RL$_{\text{ROUGE}}$) also surpasses SFT. Arguably, this improvement can be attributed to RL's enhanced exploration and generalization capabilities, which in a reference-based instantiation also help to discover more effective generation patterns and mitigate exposure bias (Paulus et al., 2018).

6.3 Results on Varying Source Documents

It is worth exploring how the number of source documents influences model performance. We display the performance across different document number groups (distribution in Appendix C) of News and XScience datasets in Figures 3 and 4, respectively. We report the geometric mean of ROUGE scores for comparison among BASE, RL$_{\text{HF}}$, SFT, and our two models: RL$_{\text{TOPIC-7B}}$ and RL$_{\text{TOPIC+ROUGE}}$.

For Multi-News, ROUGE-M scores decline as document number increases, confirming the challenge posed by multiple long documents. Although SFT achieves top performance on two-source documents, it exhibits significant instability and performance degradation with additional source documents. In contrast, our approaches exhibit more stable performance compared to all competitors.

Multi-XScience reveals a contrasting trend: encouragingly, our models steadily improve performance with increasing numbers of source documents, a trend not observed with BASE or RL-HF. Although SFT also shows improvement with more documents, its results fluctuate significantly, making it less reliable. Our RL-trained models, enhanced by topical information aligned with each source document, deliver the most consistent and superior performance, demonstrating clear advantages in practical multi-document scenarios.

6.4 Qualitative Analysis

Human Evaluation on Topic Quality To verify our hypothesis that explicitly extracted topic

	Model	Overlap-Based				Similarity-Based		Topic Alignment	
		Rouge-1	Rouge-2	Rouge-L	Rouge-M	BERT	LLM2V	CovRatio	PreRatio
News	SFT	**43.43**	**14.36**	20.76	**23.28**	.854	.815	.530	.665
	RL_{ROUGE}	41.43	12.70	19.19	21.61	.849	.802	.533	.670
	$\text{RL}_{\text{TOPIC-7B+ROUGE}}$ (ours)	43.05	13.66	**21.06**	23.14	**.856**	**.827**	**.543**	**.686**
XScience	SFT	33.21	**9.28**	18.03	17.71	.847	.749	.479	.503
	RL_{ROUGE}	35.20	8.32	**18.07**	17.43	.849	.755	.542	.543
	$\text{RL}_{\text{TOPIC-7B+ROUGE}}$ (ours)	**35.61**	8.80	18.04	**17.81**	**.851**	**.763**	**.555**	**.561**

Table 3: On Mutli-News and Multi-XScience datasets, we compare our RL-trained models with topic and rouge rewards (ours) against supervised fine-tuning (SFT) and RL-trained with solely ROUGE reward.

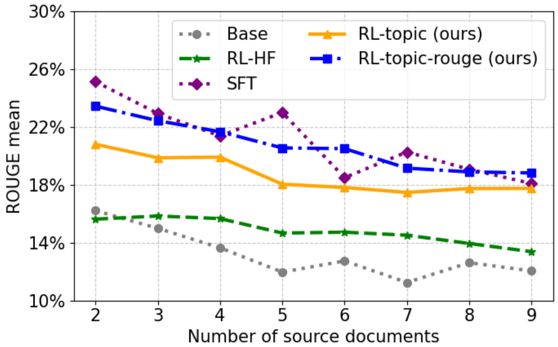


Figure 3: Model performance under different number of source document groups on Multi-News test set.

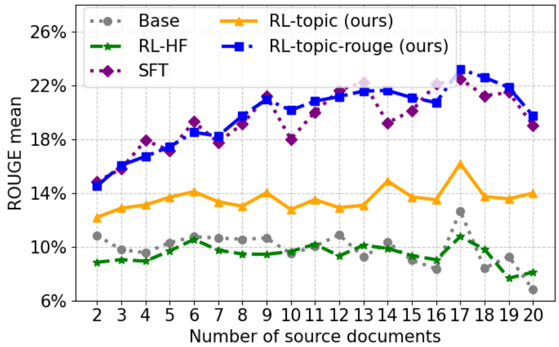


Figure 4: Model performance under different number of source document groups on Multi-XScience test set.

phrases can effectively guide MDS, we conduct a human evaluation assessing the quality of topics generated by Qwen 7B and Qwen 0.5B models. Specifically, we evaluated four criteria—*Relevance*, *Coverage*, *Specificity*, and *Redundancy*—using a 5-point Likert scale. Detailed evaluation guidelines and results are provided in Appendix D.

In brief, evaluation results indicate that the 7B model consistently produces precise and conceptually rich topic phrases, often comprising multi-word expressions. In contrast, the 0.5B model tends to generate topic that, while relevant, lack sufficient coverage and specificity, and with semantic redundancy. These findings support the benefits of *teacher-supervised* framework, where larger models with superior topic-modeling capabilities effectively guide smaller models through topic distillation, thereby improving summarization performance.

Failure Cases in Generation During evaluation, we observe that models occasionally produce excessively long and repetitive outputs at inference time. We quantify the frequency of such failure cases across all model variants (see Appendix E) and find that the SFT model is most prone to this issue, with over 3% of instances failing to generate coherent sentences. This partly accounts for the high variance observed in its performance. In contrast, the RL-trained model with human preference rewards, as well as our proposed models with topic cues, exhibit greater stability, with minimal occurrence of such degenerate outputs ($< 0.2\%$).

7 Conclusion

We introduce an interpretable, reference-free topic-guided RL approach for MDS, leveraging a novel topic-F1 reward that aligns summary topics with source documents. Integrated within the GRPO framework, our method consistently outperformed strong baselines, demonstrating the value of explicit topic guidance. Looking forward, we aim to enrich our framework by exploring advanced neural topic modeling techniques (Bianchi et al., 2021; Fang et al., 2024) for more refined topical guidance. Moreover, incorporating innovative reward signals, such as LLM-as-a-judge evaluation (Zheng et al., 2023; Liusie et al., 2024), could further align summaries with human preferences and enhance self-consistency. Extending our topic-guided approach to interactive, query-based scenarios–where users specify key points to summarize–also presents an exciting future direction.

Limitations

In our experiments, we focus primarily on models from the Qwen series, selected for their strong performance across diverse NLP tasks and the availability of multiple model sizes. This choice enables us to highlight and isolate the impact of topic alignment, avoiding potential confounding factors from different architectures across different LLM families. Furthermore, the policy model employed in our current setup is a 0.5B parameter model. Though exploring larger models is promising, the substantial computational cost limits such experiments in the current study. Nonetheless, our results clearly demonstrate the effectiveness of the topic reward approach even with this modestly sized model, laying a solid foundation for future studies that may scale to more powerful models.

Evaluating text summarization continues to be challenging due to the multifaceted aspects involved in assessing summary quality (Kryscinski et al., 2019; Fabbri et al., 2021; Goyal et al., 2022). In our work, we employ a range of automatic metrics, including traditional methods (ROUGE), embedding-based approaches, and our newly proposed topic alignment metrics, which notably do not require reference summaries.

Although we acknowledge the availability of other reference-free metrics, integrating them effectively into our task–summarization of multiple lengthy source documents–is nontrivial. For example, our preliminary analysis with an entailment-based metric revealed that factual scores assigned to gold-standard summaries were sometimes lower than those assigned to zero-shot prompted summaries. Upon careful inspection, we discovered this occurred because certain prompted summaries heavily mirrored the first paragraph of source texts, resulting in disproportionately high entailment scores at the sentence level. This scoring pattern, however, does not accurately reflect comprehensive summary quality, as a good summary must synthesize information distributed across multiple documents. Thus, we propose our topic coverage and precision scores as a more balanced evaluation approach tailored for this task.

Ethical Statement

We have taken proactive steps to address ethical concerns related to our research. Our corpora were carefully selected to minimize potential issues with biased or hateful content. For human evaluation, we clearly instructed annotators to remain vigilant and identify any biased or inappropriate language within the data. The annotators participated voluntarily without specific compensation; however, they were encouraged to use the results of their evaluation work for their academic studies.

References

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. OpenAsp: A benchmark for multi-document open aspect-based summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1991, Singapore. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Catarina G Belem, Pouya Pezeshkpour, Hayate Iso, Seiji Maekawa, Nikita Bhutani, and Estevam Hruschka. 2024. From single to multi: How llms hallucinate in multi-document summarization. *arXiv preprint arXiv:2410.13961*.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.

Peng Cui and Le Hu. 2021. Topic-guided abstractive multi-document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1463–1472.

Peng Cui, Le Hu, and Yuanchao Liu. 2020. Enhancing extractive text summarization with topic-aware graph neural networks. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander Richard Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

Hugging Face. 2025. Open r1: A fully open reproduction of deepseek-r1.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Gao, Christian Meyer, Mohsen Mesgar, and Iryna Gurevych. 2019. Reward learning for efficient reinforcement learning in extractive document summarisation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2350–2356.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Gili Lior, Avi Caciularu, Arie Cattan, Shahar Levy, Ori Shapira, and Gabriel Stanovsky. 2024. Seam: A stochastic benchmark for multi-document tasks. *arXiv preprint arXiv:2406.16086*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081.

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages

4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Ani Nenkova and Lucy Vanderwende. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

You Ouyang, Sujian Li, and Wenjie Li. 2007. Developing learning strategies for topic-based summarization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 79–86.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779, Online. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

DM Roijers, P Vamplew, S Whiteson, and R Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Harm Van Seijen, Mehdi Fatemi, Joshua Romoff, Romain Laroche, Tavian Barnes, and Jeffrey Tsang. 2017. Hybrid reward architecture for reinforcement learning. *Advances in neural information processing systems*, 30.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Noah Wang, Feiyu Duan, Yibo Zhang, Wangchunshu Zhou, Ke Xu, Wenhao Huang, and Jie Fu. 2024. Positionid: Llms can control lengths, copy and paste with explicit positional awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16877–16915.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 485–497.

Yang Wei. 2012. Document summarization method based on heterogeneous graph. In *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, pages 1285–1289. IEEE.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Ming Zhang, Jiyu Lu, Jiahao Yang, Jun Zhou, Meilin Wan, and Xuejun Zhang. 2024. From coarse to fine: Enhancing multi-document summarization with multi-granularity relationship-based extractor. *Information Processing & Management*, 61(3):103696.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 46595–46623.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and 1 others. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Prompt Template for MDS and Topic Modeling

We present various prompts used in our work, both in zero-shot setting (Table 4), and RL training (Table 5, 6, and 7):

MDS with Topic Labels Prompt in Zero-shot

A conversation between User and Assistant. The user provides news articles and topic labels, and the Assistant produces a short summary. The summary contains no more than **ten sentences** and **only** based on information from the provided articles and topic labels.
Documents: `{Doc 1 text}`, `{Doc 1 topics}`, `{Doc 2 text}`, `{Doc 2 topics}`, `...`, `{Doc K text}`, `{Doc K topics}`.
Assistant:

Table 4: Prompt template used by the Qwen2.5-0.5B / 1.5B / 7B in preliminary zero-shot experiments to test the performance with integrated topic labels (See §3).

MDS Prompt in RL training (Multi-News)

A conversation between User and Assistant. The user provides news articles, and the Assistant produces a short summary. The summary contains no more than **ten sentences** and **only** based on information from the provided articles.
Documents: `{Doc 1 text}`, `{Doc 2 text}`, `...`, `{Doc K text}`
Assistant:

Table 5: Prompt template used by Qwen2.5-0.5B to generate summary (rollout) from Multi-News during RL training.

MDS Prompt in RL training (Multi-XScience)

The user provides scientific articles, and the Assistant generates a related work paragraph based on the query paper's abstract and the abstracts of its referenced papers. The answer includes citations for all referenced papers (@cite_id) and be approximately **five sentences long**.
Documents: `{Doc 1 text}`, `{Doc 2 text}`, `...`, `{Doc K text}`
Assistant:

Table 6: Prompt template used by Qwen2.5-0.5B to generate summary (rollout) from Multi-XScience during RL training.

B Training Details

Hyper-Parameters Table 8 lists the hyper-parameters for GRPO training. Our training is

> **Topic Modeling Prompt in RL training**
>
> A conversation between User and Assistant. The user provides a news article, and the Assistant produces **five** key words or phrases as **topic labels**. The answer should be in the form of a list, with each item separated by a comma. Do not give any explanation or additional information.
> Document: `{Doc text}`
> Assistant:

Table 7: Prompt template used by the Qwen2.5-7B model to extract topic labels from the generated summaries (during reward calculation). Note that we pre-extract and store topic labels from the source documents before training, thus avoiding redundant topic extraction computations during the training process.

GRPO Hyperparameters	Value
Training epoch	2
Number of processes	6
Max prompt length	8092
Max completion length	1024
Gradient accumulation steps	21
Number of generations	8
Per device train batch size	4
Learning rate	$1e-6$
KL Coefficient	0.04
Epsilon	0.2
Warm-up ratio	0.1
Temperature	0.7

Table 8: Hyperparameters for GRPO training.

based on TRL (von Werra et al., 2020), adapted to our datasets and compute constraints. The best-performing checkpoint is selected based on validation reward improvements.

Due to the significant computational demands of our experiments, extensive hyperparameter optimization was impractical. Instead, we conducted pilot small-scale tests, as described in the methods section (§4), to inform our experimental setup.

For example, we observed that GRPO training is highly sensitive to **learning rate** adjustments. Although previous literature suggests using moderately higher temperatures to facilitate exploration, we found that temperatures of $1e-5$ or higher caused considerable fluctuations during training, leading to gradient explosions. Consequently, we maintained a low learning rate of $1e-6$.

Another important observation relates to the **number of completions** per input sample. We noted that increasing the number of generated samples per input improved performance, aligning with findings reported in Open-r1 (Face, 2025). However, due to the long input length, increasing the number of completions further required enlarging the training batch size, resulting in out-of-memory (OOM) errors. Therefore, we selected a sample size of 8, balancing performance gains and computational constraints.

Implementation with RLHF Reward The RL$_{\text{HUMAN-FEEDBACK}}$ reward model is designed to predict a preference score between generated answers given a specific question. During our implementation, we observed that including full source documents and generated summary as input significantly increased the computation time for calculating preference scores ($>$ 20 seconds per data point), rendering it impractical for RL training.

To address this issue, we utilized key topic phrases as a concise proxy for the original documents. This approach substantially reduced computation time while effectively preserving relevant source content. This experience also highlights that using RLHF trained rewards directly with lengthy documents is computationally prohibitive, whereas our proposed method introduces minimal computational overhead for evaluating topic alignment.

C Statistics of Number of Source Documents

As shown in Table 9: Multi-News primarily features two-source documents, while Multi-XScience has a more balanced distribution, with 2-, 3-, and 4-source inputs together making up 52% of test set.

D Human Evaluation on Topic Quality

We conducted a human evaluation to assess the quality of topic phrases generated by the Qwen2.5-7B and Qwen2.5-0.5B models. Two graduate-level annotators independently evaluated the outputs for ten randomly selected documents from the Multi-XScience training set. The evaluation was guided by four criteria–Relevance, Coverage, Specificity, and Redundancy–rated on a 5-point Likert scale (1 = poor, 5 = excellent). The detailed annotation instructions are provided in Table 10.

Annotators were first asked to read the source documents and were free to highlight or mark key phrases they considered important. Subsequently, they were presented with anonymized and randomly ordered topic lists generated by the two models. For each list, annotators assigned scores based on the four evaluation criteria.

The aggregated evaluation results for twenty

Datasets	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
News	54.6	27.8	10.9	3.9	1.7	0.7	0.2	0.2	-	-	-	-	-	-	-	-	-	-	-
XScience	23.9	13.3	15.5	10.6	6.6	6.0	4.3	2.9	4.1	2.9	2.9	1.8	1.4	1.3	0.7	0.6	0.6	0.3	0.3

Table 9: Distribution of the number of source documents in Multi-News (News) and Multi-XScience (Science) datasets, both in the test subset.

Criterion	Guiding Question
Relevance	Do the phrases reflect the central themes or key ideas of the document?
Coverage	Do the phrases collectively represent diverse and important parts?
Specificity	Are the phrases informative and precise, not vague or overly general?
Redundancy	Are any phrases repeated or semantically overlapping?

Table 10: Evaluation criteria instructions.

Model	Relevance	Coverage	Specificity	Redundancy
7B	5.0	5.0	5.0	5.0
0.5B	3.8	2.8	2.4	3.2

Table 11: Model topic evaluation summary.

topic sets (ten documents, two models) are summarized in Table 11. The findings indicate that Qwen2.5-7B consistently outperforms Qwen2.5-0.5B, producing more precise, accurate, and comprehensive topic phrases. In contrast, the 0.5B model exhibits notable deficiencies in coverage and specificity. Table 12 presents several representative examples to illustrate the qualitative differences between the models. In these examples, topics highlighted in red were identified by annotators as inappropriate, typically due to being overly generic or lacking clarity.

E Qualitative Results

Table 13 presents the percentage of failure cases in which the model produces excessively long and incoherent outputs. Notably, the SFT model exhibits the highest failure rate, whereas models trained with reinforcement learning show the lowest. Below, we provide an example of a failure case generated by the SFT model, corresponding to test example #21 in the Multi-XScience dataset:

“ *In recent years, many new methods have been developed to solve the blind image denoising problem. First, the mixture of Gaussian distribution @cite_21 @cite_8 @cite_13 @cite_30 @cite_40 @cite_10 @cite_19 @cite_23 @cite_9 @cite_32 @cite_6 @cite_25 @cite_18 @cite_23 @cite_32 @cite_18 @cite_18 @cite_13 ...*” – the model continues repeating content until it exhausts the maximum output length defined by vLLM.

Qwen-7B topics	Qwen-0.5B topics
Communication strategies, collaborative problem solving, resource limitations, task requirements, experimental simulations	Effective, problem solving, resource bounded, communication, collaborative
Principle of Parsimony, Task-Oriented Dialogue, Recovery Strategies, Information Transfer, HCRC Map Task	Parsimonious, task-oriented, information, recovery, dialogue
Automatic Text Categorization, WordNet, Vector Space Model, Rocchio Algorithm, Widrow-Hoff Algorithm	WordNet, Rocchio, Widrow-Hoff, category, low frequency
Natural Language Processing, TextTiling, TileBars, Cougar, Topic Labeling	Contextual, text, topic, retrieval, display
Text categorization, WordNet, lexical databases, training collections, performance comparison	Auto text categorization, lexical databases, training collections, WordNet, WordNet-based

Table 12: Examples of generated topics from Qwen2.5-7B and Qwen2.5-0.5B models. Topics highlighted in red are considered as inappropriate topics.

Model	Multi-News (%)	Multi-XScience (%)
BASE	1.37	0.04
BASE$_{\text{TOPIC}}$	2.41	0.14
SFT	3.92	3.14
RL$_{\text{HF}}$	0.07	0.00
RL$_{\text{ROUGE}}$	0.52	0.02
RL$_{\text{TOPIC-0.5B}}$ (ours)	0.18	0.00
RL$_{\text{TOPIC-7B}}$ (ours)	0.12	0.00
RL$_{\text{TOPIC+ROUGE}}$ (ours)	0.14	0.08

Table 13: Percentage of failure cases where model generates repetitive and long output (e.g., $> 2,500$ tokens).