30th ANNIVERSARY EMNLP 2025 Suzhou, China | 中国苏州

Paper Link Code Link

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

Chuyuan Li*, Austin Xu^, Shafiq Joty^, Giuseppe Carenini*

*Department of Computer Science, University of British Columbia, ^Salesforce AI Research

chuyuan.li@ubc.ca, carenini@cs.ubc.ca, {austin.xu, sjoty}@salesforce.com

Background & Motivation

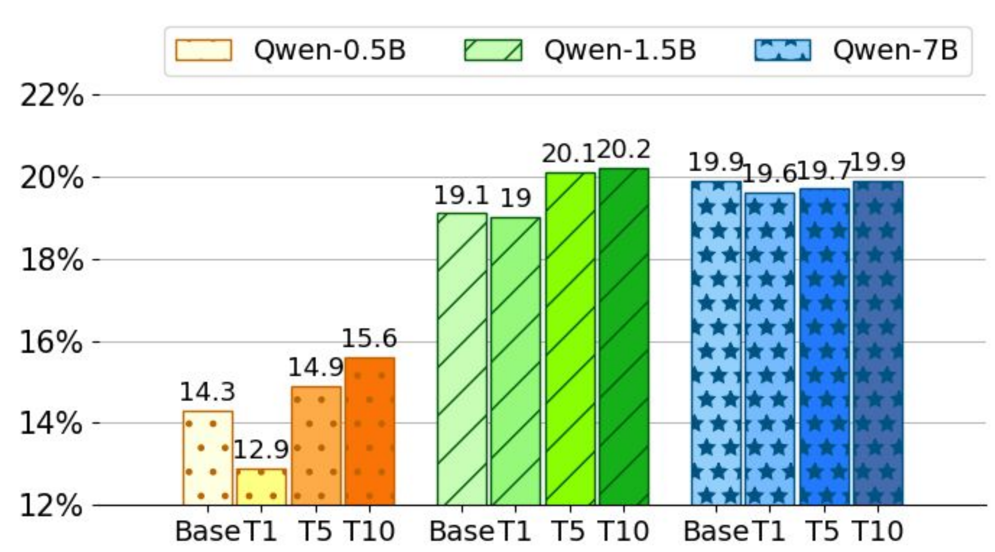
- **Multi-Document Summarization (MDS)**: Challenges in integrating multiple sources and maintaining coherence and topical relevance.
- **Large Language Models (LLMs)**: Impressive results in single-document summarization, but need to improve on content relevance, coherence, and topic consistency with MDS.
- **Proposal**:
  - i. Incorporation of high-level discourse information to guide MDS → Topics offer a **global discourse structure**
  - ii. Explicit usage of topic labels in MDS → **Direct prompting with topic labels**
  - iii. Injection of topic awareness into training objective → **Reinforcement learning (GRPO) with topic-guided reward**

Direct Prompting

- Prompting with topics:

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

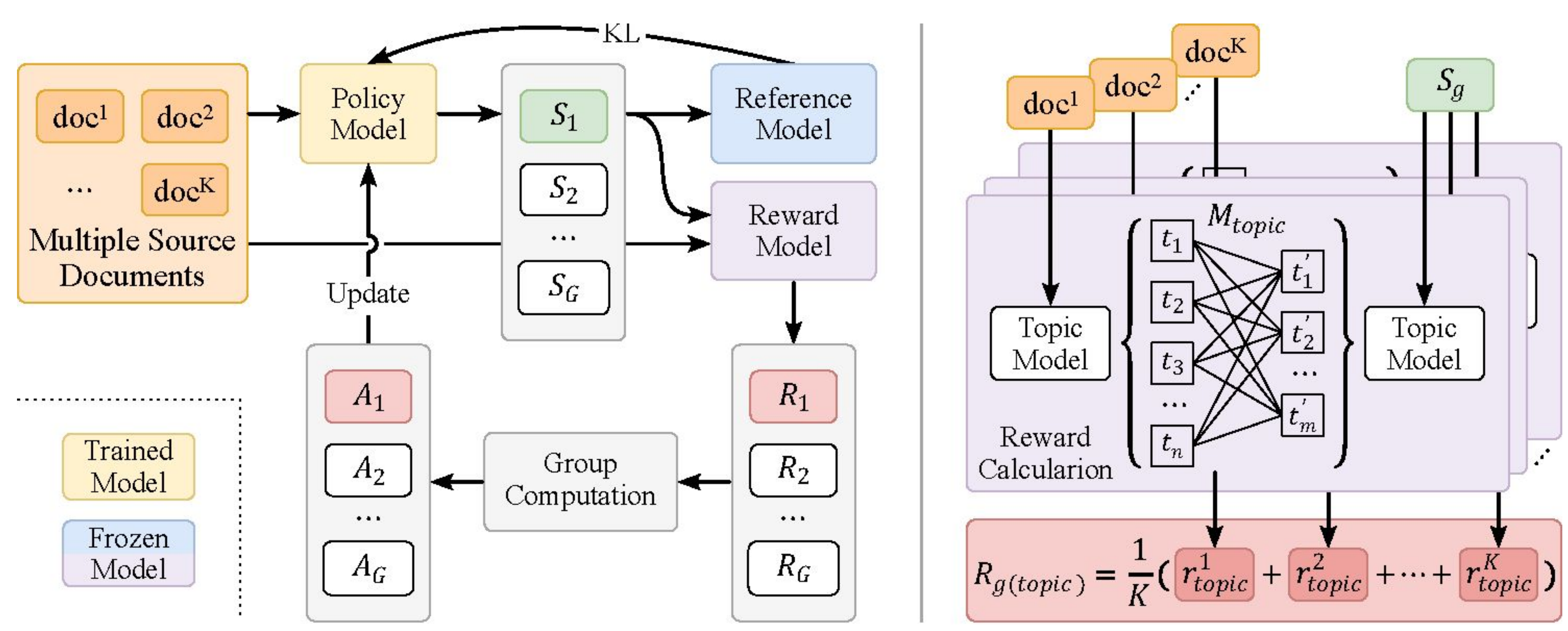
- *Teacher-supervision* mode: larger LLM Qwen2.5-7B provides topic labels to "student" LLMs 0.5B and 1.5B
- Varying number of topic labels:
  - T={1, 5, 10}



→ Smaller base models (0.5B, 1.5B) benefit from improved topic information.
→ 7B model itself does not show gains from self-generated topic labels.
→ 1 label: overly constraints summ; more labels (T5, T10) show benefits.

LLM RL with Topic-guided Reward

- Topic-guided reward: **Topic-F1**
  - Construction of similarity matrix M_topic, where M_ij represent cosine similarity b/t topic embeddings of a pair of topic phrases from [source_doc, generated summary]
  - Reward calculation r_topic from M_topic:
    - Coverage: avg max similarity b/t each source topic and its most similar summary topic
    - Precision: avg max similarity b/t each summary topic and its most similar source topic



(a) GRPO Training (b) Topic-Guided Reward

$$R_{g(topic)} = \frac{1}{K}\left(r^1_{topic} + r^2_{topic} + \cdots + r^K_{topic}\right)$$

- Length-penalty reward (token-level): $R_{\text{len}} = \exp\left(-\frac{|L_{\text{exp}} - L_{\text{sum}}|}{L_{\text{exp}}}\right)$
- Reward weighting: Inverse std.dev weighting, emphasis factor $w_r^{\text{norm}} = \frac{w_r \times \text{factor}_r}{\sum_k (w_k \times \text{factor}_k)}$
- GRPO training: Advantage estimation

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

Experiments

- Dataset: Multi-News, Multi-XScience
- Evaluation: overlap, similarity, topic-align
- Model comparisons
  - RL, topic-reward (ours):
    - Policy model: Qwen2.5-0.5B
    - Reward model: Qwen2.5-0.5B, 7B
  - RL, human-feedback:
    - Reward model: deberta-v3-large-v2
  - RL, rouge-reward (reference-based)
    - Further combined with our topic-reward
  - Base (no RL): Qwen2.5-0.5B, 7B
  - SFT: Qwen2.5-0.5B

Further investigation (§6.3, 6.4, 6.5, 6.6)

- LLM-as-a-judge evaluation
  - Judge: GPT-4.1
  - (multiple) pairwise comparisons: ours is consistently the winner
- Human evaluation on topic quality
  - *Relevance, cov., specificity, redundancy*
  - 7B model produces precise and rich topics
- Analysis on varying N of source documents
  - Topic-RL model most stable
- RL combined with *Best-of-n* strategy (inference time scaling)
  ⇒ RL+scaling > RL > Base + scaling > Base.

	Model	IM	RM	Overlap-Based Rouge-1	Overlap-Based Rouge-2	Overlap-Based Rouge-L	Overlap-Based Rouge-M	Similarity-Based BERT	Similarity-Based LLM2V	Topic Alignment CovRatio	Topic Alignment PreRatio
Reference-free methods											
News	Base (0.5B)	0.5B	-	27.22	7.28	15.03	14.31	.842	.721	.513	.622
	Base (7B)	7B	-	37.09	10.77	**19.77**	19.91	**.845**	.796	.538	.672
	$\text{Base}_{\text{Topic-7B}}$	0.5B	7B	28.62	8.60	15.83	15.73	.844	.733	.521	.632
	$\text{RL}_{\text{human-feedback}}$	0.5B	0.3B	33.07	6.99	17.29	15.58	.819	.706	.492	.583
	$\text{RL}_{\text{Topic-0.5B}}$ (ours)	0.5B	0.5B	38.63	10.72	18.81	19.82	**.845**	.793	.536	.672
	$\text{RL}_{\text{Topic-7B}}$ (ours)	0.5B	7B	**39.62**	**10.97**	18.97	**20.20**	**.845**	**.798**	**.540**	**.676**
XScience	Base (0.5B)	0.5B	-	25.05	4.16	13.47	11.19	.822	.637	.490	.480
	Base (7B)	7B	-	30.08	5.06	15.31	13.26	.838	.728	.550	.549
	$\text{Base}_{\text{Topic-7B}}$	0.5B	7B	25.62	4.09	13.93	11.34	.828	.655	.482	.479
	$\text{RL}_{\text{human-feedback}}$	0.5B	0.3B	26.78	2.90	13.87	10.25	.832	.622	.506	.507
	$\text{RL}_{\text{Topic-0.5B}}$ (ours)	0.5B	0.5B	29.47	4.79	15.90	13.09	.835	.721	.548	.549
	$\text{RL}_{\text{Topic-7B}}$ (ours)	0.5B	7B	**30.45**	**5.38**	**16.26**	**13.86**	**.847**	**.741**	**.554**	**.560**
Reference-based methods											
News	SFT	0.5B	-	43.24	14.28	20.51	23.18	.852	.813	.529	.665
	RL_{Rouge}	0.5B	0.5B	41.43	12.70	19.19	21.61	.849	.802	.533	.670
	$\text{RL}_{\text{Topic-7B+Rouge}}$ (ours)	0.5B	7B	**43.51**	**14.31**	**21.55***	**23.40**	**.857***	**.823***	**.543***	**.683***
XSci	SFT	0.5B	-	33.61	**9.25**	**18.28**	**17.72**	.850	.750	.480	.510
	RL_{Rouge}	0.5B	0.5B	35.20	8.32	18.07	17.43	.849	.755	.542	.543
	$\text{RL}_{\text{Topic-7B+Rouge}}$ (ours)	0.5B	7B	**36.16***	8.96	18.15	17.71	**.852***	**.765***	**.557***	**.569***

	Model	Overlap-Based Rouge-1	Overlap-Based Rouge-2	Overlap-Based Rouge-L	Overlap-Based Rouge-M	Similarity-Based BERT	Similarity-Based LLM2V	Topic Alignment CovRatio	Topic Alignment PreRatio	Topic Alignment F1
News	Base (0.5B)	27.22	7.28	15.03	14.31	.842	.721	.513	.622	.562
	Base (0.5B) + *best-of-n*	29.27	8.68	15.87	15.92	**.847**	.738	.517	.647	.575
	$\text{RL}_{\text{Topic-7B+Rouge}}$	39.62	10.97	18.97	20.20	.845	.798	.540	.676	.600
	$\text{RL}_{\text{Topic-7B+Rouge}}$ + *best-of-n*	**40.95**	**12.03**	**19.63**	**21.30**	.842	**.798**	**.546**	**.683**	**.607**
XScience	Base (0.5B)	25.05	4.16	13.47	11.19	.822	.637	.490	.480	.485
	Base (0.5B) + *best-of-n*	27.88	4.64	14.68	12.38	.831	.708	.523	.518	.521
	$\text{RL}_{\text{Topic-7B}}$	30.45	5.38	16.26	13.86	.847	.741	.554	.560	.557
	$\text{RL}_{\text{Topic-7B}}$ + *best-of-n*	**30.94**	**5.55**	**16.37**	**14.11**	**.849**	**.753**	**.562**	**.579**	**.570**