Large Language Model-Powered Query-Driven Event Timeline Summarization in Industrial Search

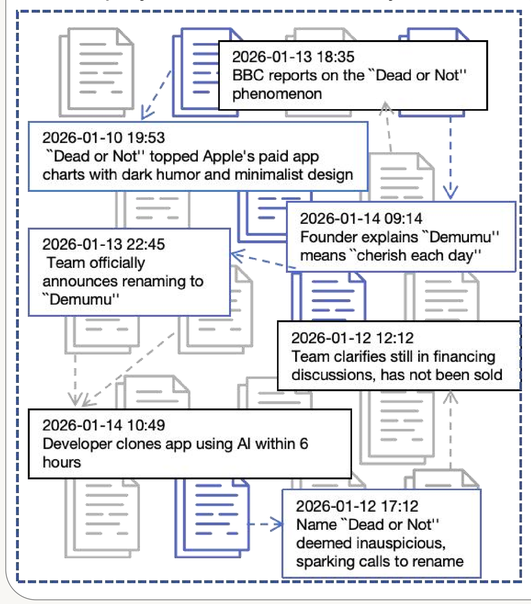
Mingyue Wang, Xingyu Xie, Hang Yang, Li Gao*, Lixin Su, Ge Chen, Dawei Yin, Daiting Shi

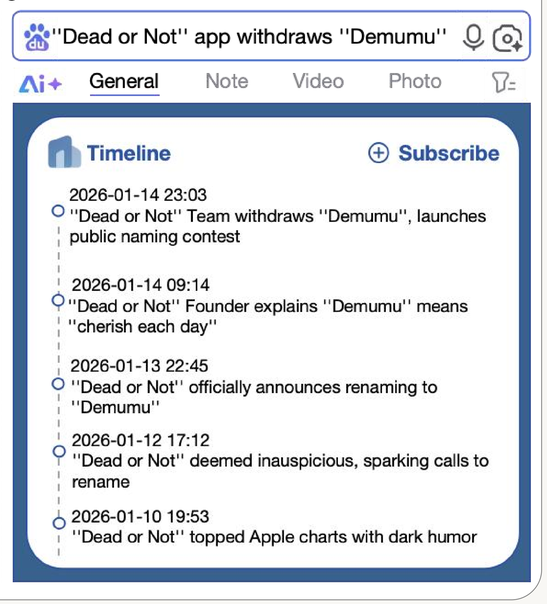
KDD 2026 August 9–13, 2026 Jeju, Korea

1 Why Query-driven Timelines?

Users do not want a broad topic digest. They want the story behind one event: what happened first, what caused it, and what came next.

- Millions of documents arrive every day, but only a few events explain the query.
- Retrieved candidates are noisy: core events, peripheral coverage, duplicates, and distractors.
- Display summaries must stay short enough for search cards: 5-15 Chinese characters.





2 Problem Setup

Task Definition

Given a query event e_q and a large-scale event database constructed from streaming documents, our goal is to **generate a query-driven event timeline:**

$$T =< e_1, e_2, \ldots, e_q >$$

that explains how a queried event evolves by selecting the most relevant events, organizing them into a coherent narrative, and satisfying industrial deployment constraints.

Challenge 1: Noisy Candidate Events

Pro: Retrieved candidate events contain substantial noise, making it difficult to identify the events that truly explain the queried event.

Sol: Formulate timeline generation as a query-driven reasoning task and use a multi-task fine-tuned LLM to filter noisy events through temporal and causal reasoning.

Challenge 2: Temporal & Causal Reasoning

Pro: Semantic similarity alone cannot produce coherent timelines; the model must infer temporal order, causal dependencies, and narrative completeness.

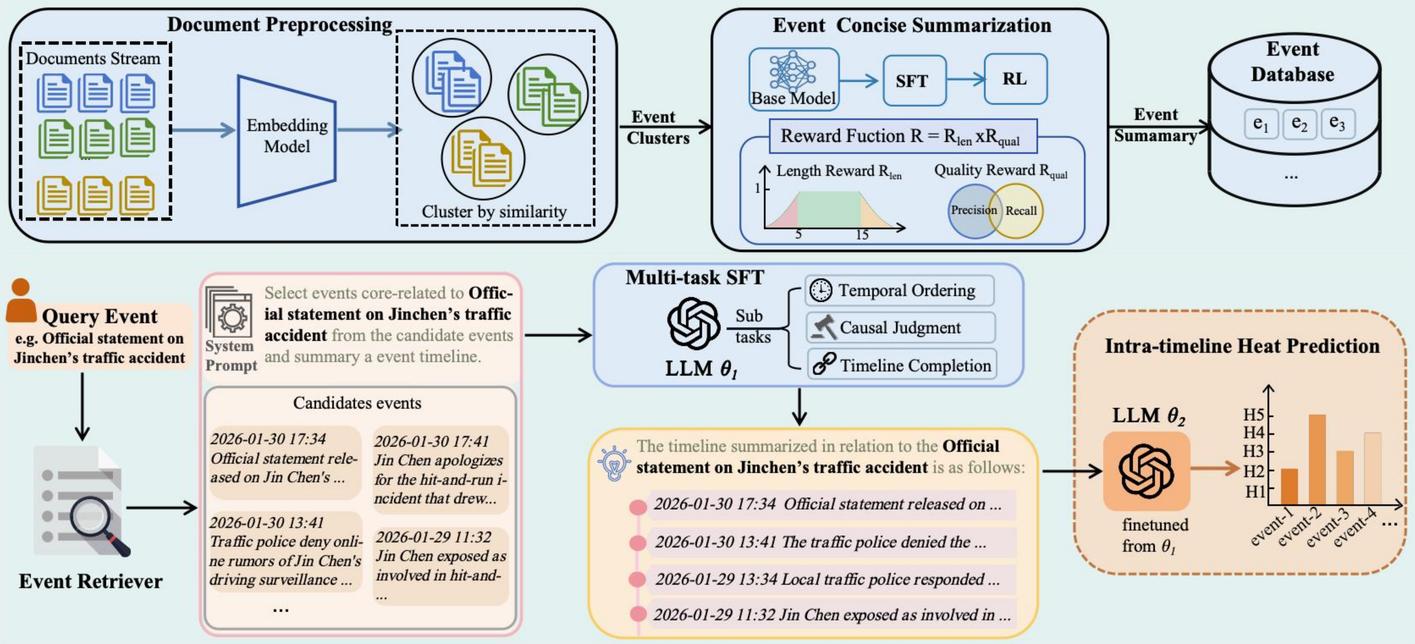
Sol: Introduce three auxiliary tasks—Temporal Ordering, Causal Judgment, and Timeline Completion—to explicitly enhance reasoning over event relationships.

Challenge 3: Industrial Deployment Constraints

Pro: Industrial search requires concise, informative event summaries that satisfy strict display constraints while supporting efficient online serving.

Sol: Develop an RL-based Event Concise Summarization module to jointly optimize semantic quality and length compliance for production-ready deployment.

3 Our Method



4 Results

✨ Main Performance

Our 7B model achieves performance comparable to DeepSeek-R1-671B using only ~1% parameters.

Base Model	P↑	R↑	F1↑	AR-1↑	τ ↑
Qwen2.5-3B-Instruct (0-shot)†	59.2	60.1	59.6	59.3	0.65
Qwen2.5-3B-Instruct (5-shot)†	60.5	61.8	61.1	60.5	0.67
Qwen2.5-7B-Instruct (0-shot)†	62.7	64.8	63.7	63.1	0.68
Qwen2.5-7B-Instruct (5-shot)†	64.1	66.2	65.1	64.5	0.70
DeepSeek-R1-671B (0-shot)†	73.6	78.7	76.1	74.5	0.89
DeepSeek-R1-671B (5-shot)†	73.9	79.2	76.4	74.8	0.89
Qwen2.5-3B-Instruct (fine-tuned)	66.8	74.5	70.4	68.1	0.83
Qwen2.5-7B-Instruct (fine-tuned)	**73.2**	**79.5**	**76.2**	**73.8**	**0.87**

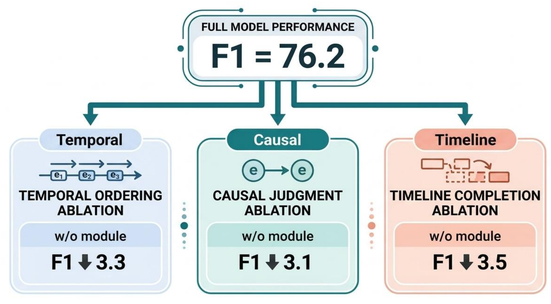
✨ RL Event Summary

RL significantly improves both semantic fidelity and length control, producing concise, production-ready event summaries for large-scale search deployment.

Method	R-1↑	R-2↑	R-L↑	Len↑
DeepSeek-R1-671B (zero-shot)	66.8	53.9	60.2	76.8
DeepSeek-R1-671B (few-shot)	69.5	57.2	63.8	80.5
Qwen2.5-3B-Instruct (SFT)	66.3	53.7	59.8	78.3
Qwen2.5-7B-Instruct (SFT)	68.1	55.4	61.5	83.5
Qwen2.5-3B-Instruct (SFT+RL)	68.3	56.2	60.6	85.5
Qwen2.5-7B-Instruct (SFT+RL)	**71.9**	**59.8**	**65.8**	**88.2**

✨ Contribution of Auxiliary Tasks

Every auxiliary task contributes to timeline reasoning.



✨ Online Deployment

Controlled online A/B tests on Baidu Search, covering 100+ event timelines and millions of user interactions, demonstrate significant improvements.

+5.5%	+4.6%	+4.4%
CTR	dwell time	exploration depth

Contact
wangmy.mia@gmail.com
gaoli.sinh@gmail.com

arXiv: 2605.27066