Prompteam at UZH Shared Task 2026: RAG-Augmented Classification and Cosine-Filtered Relation Prediction for UN Resolutions

Siddhartha Khandelwal
GLA University
Mathura, UP, India
siddharthakhandelwal9@gmail.com

Jyotsana Bhardwaj
GLA University
Mathura, UP, India
jyotsanab15@gmail.com

Abstract

We describe our system for the UZH ArgMining 2026 Shared Task on reconstructing argumentative structure in UN/UNESCO resolutions. The task requires (1) classifying paragraph types and assigning thematic tags from a 141-label taxonomy, and (2) predicting directed argumentative relations between paragraphs. Our pipeline combines a quantised Qwen2.5-7B-Instruct model with retrieval-augmented generation (RAG) backed by FAISS-indexed dense embeddings for few-shot prompting and tag candidate pre-filtering. For relation prediction, we apply a sliding-window cosine pre-filter that reduces the quadratic pair space to near-linear cost. A parallelisable, fault-tolerant pipeline with atomic checkpointing enabled complete processing of 2,959 paragraphs across three concurrent Kaggle T4 sessions despite 12-hour GPU limits. Our system achieved **2nd place** overall on the shared task leaderboard.

1 Introduction

The UZH ArgMining 2026 Shared Task (Gao and others, 2026) targets the analysis of argumentative structure in UN/UNESCO resolutions. **Subtask 1** requires per-paragraph prediction of (a) a binary structural type (*preambular* vs. *operative*) and (b) a multi-label subset from a 141-label thematic taxonomy spanning 15 dimensions. **Subtask 2** requires identifying directed relations between paragraph pairs from four types: *supporting*, *contradictive*, *complemental*, and *modifying*. The final ranking averages an automated F1 score and an LLM-as-a-Judge score (0–100) that rates chain-of-thought reasoning quality. All teams must use open-weight models with $\leq$8B parameters.

Our contributions are:

- A RAG-based few-shot classification pipeline using FAISS-indexed dense embeddings for tag candidate pre-filtering and dynamic example retrieval (§3.3).
- A cosine-similarity pre-filter reducing relation prediction from $O(N^2)$ to near-linear cost (§3.4).
- A fault-tolerant, parallelisable design with atomic checkpointing across Kaggle's 12-hour GPU limit (§3.5).
- **2nd place** on the shared task leaderboard.

2 Related Work

Argument mining has progressed from essay-level parsing (Stab and Gurevych, 2017) and MST-based discourse prediction (Peldszus and Stede, 2015) to large-scale political text analysis (Lawrence and Reed, 2020). The UN-RES corpus (Gao and others, 2026) extends argument mining to multilingual UN resolutions with fine-grained thematic tagging. RAG pipelines (Lewis et al., 2020) ground LLM predictions in retrieved evidence; dynamically selecting semantically similar demonstrations outperforms random sampling in few-shot settings (Rubin et al., 2022). We apply this paradigm jointly to tag candidate retrieval and few-shot example selection, and reduce the quadratic cost of pairwise relation prediction via embedding-based cosine pre-filtering.

3 System Description

3.1 Data

The shared task provides both a training set and a held-out test set in JSON format (Gao and others, 2026). The **training data** consists of 2,695 parsed UN resolutions as raw text in French, drawn from the UN-RES dataset (Gao et al., 2025), with machine-generated English translations produced using Helsinki-NLP/opus-mt-fr-en. Use of the training data is unrestricted; the organisers encourage techniques with strong LLM reasoning focus

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 136–139
July 3, 2026 ©2026 Association for Computational Linguistics

(e.g. RAG, in-context learning). The **test data** comprises 45 parsed documents (resolutions and recommendations) from the UNESCO International Bureau of Education's International Conference on Education (1934–2008), each containing up to three resolutions, annotated at paragraph level in French (with English translations generated using gpt-4.1-mini). In total, the test set contains 89 individual resolutions spanning 2,959 paragraphs. Paragraph-level annotations include a binary structural type (*preambular* vs. *operative*), multi-label thematic tags from a 141-label taxonomy across 15 dimensions (provided in a CSV file), and directed inter-paragraph relations from four categories: *supporting*, *contradictive*, *complemental*, and *modifying*.

3.2 Model Stack

Our LLM is `Qwen2.5-7B-Instruct` (Qwen Team, 2025) in INT8 quantisation (Dettmers et al., 2022) (∼7 GB VRAM). Sentence embeddings use `all-mpnet-base-v2` (768-d) (Reimers and Gurevych, 2019) with FAISS `IndexFlatIP` (Johnson et al., 2021) for exact cosine search.

3.3 Subtask 1: Classification

Each paragraph passes through a four-stage pipeline (Figure 1): (1) **Embed** the paragraph into a 768-d vector; (2) **Tag retrieval** via cosine search returns the top-20 candidate tags; (3) **RAG few-shot** retrieves the 5 most similar annotated paragraphs as demonstrations; (4) **LLM classify** with a structured prompt yields a JSON with fields `think` (chain-of-thought), `type`, and `tags`. A heuristic fallback (first-word rule) applies if all three JSON-parse retries fail; output tags are sanitised against the taxonomy.

Figure 1: Subtask 1 pipeline per paragraph.

Prompt design. The system prompt defines the two paragraph types with canonical opening keywords, tag selection rules emphasising precision, and a mandatory `think` template with four reasoning steps (Wei et al., 2022): identify keyword → decide type → evaluate candidates → justify tags. Few-shot examples are ordered by descending similarity (Rubin et al., 2022).

3.4 Subtask 2: Relation Prediction

For each paragraph B, we determine which earlier paragraph A shares an argumentative link. Two strategies reduce the $O(N^2)$ cost (Figure 2):

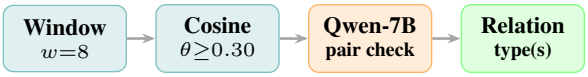

Figure 2: Subtask 2 pipeline per paragraph pair.

Sliding window. Paragraph B is compared only against the preceding w=8 paragraphs, exploiting sequential rhetorical structure. **Cosine pre-filter.** Only pairs with cosine similarity $\geq\theta$=0.30 reach the LLM, eliminating ∼65% of candidates. The prompt requests four-step reasoning: summarise A → summarise B → compare themes → assign relation(s).

3.5 Engineering Design

Parallel chunking. Both tasks distribute work across N=3 concurrent Kaggle notebooks via modulo indexing ($i \bmod N = K$), with a merge step deduplicating by composite key (`doc_id||para_id`). Full hyperparameter settings are provided in Appendix A.

Atomic checkpointing. Results are written to a `.tmp` file and atomically renamed, preventing corruption on session termination. On resume, processed entries are skipped.

4 Results

Our system achieved **2nd place** on the ArgMining 2026 leaderboard.

4.1 Subtask 1: Classification

Metric	Value
Total paragraphs	2,959
Documents	89
Duplicates	0
Preambular / Operative	58.3% / 41.7%
Avg. tags per paragraph	2.4
Avg. `think` length	74 words

Table 1: Subtask 1 coverage and distribution.

Table 1 shows full coverage of all 2,959 paragraphs with zero duplicates after the composite-key fix. The type split (58.3% preambular) is consistent with UN resolution structure. The average of 2.4 tags per paragraph reflects focused multi-label predictions; the 74-word `think` field provides sufficient reasoning depth for LLM-as-a-Judge scoring.

4.2 Subtask 2: Relations

Relation Type	Share
Supporting	44%
Complemental	31%
Modifying	18%
Contradictive	7%
Pairs eliminated by pre-filter	∼65%

Table 2: Subtask 2 relation distribution.

Table 2 shows that *supporting* relations dominate (44%), as expected in consensus-driven UN documents. *Contradictive* relations are rare (7%). The cosine pre-filter eliminated ∼65% of candidate pairs, substantially reducing LLM inference cost.

4.3 Error Analysis

Two recurring failure modes were identified. First, with only k=20 tag candidates pre-filtered per paragraph, rare labels (e.g., `INFRA_WASH`) may never enter the candidate set. Second, the cosine threshold occasionally filters out argumentatively related pairs with dissimilar vocabulary, particularly for *contradictive* relations.

5 Conclusion

We presented a complete, fault-tolerant pipeline for the UZH ArgMining 2026 Shared Task, achieving 2nd place by combining a quantised 7B instruction-tuned model with FAISS-backed retrieval-augmented generation and cosine-pre-filtered relation prediction. Key takeaways: (i) composite keys are essential for paragraph-level checkpointing across multi-document corpora; (ii) cosine pre-filtering reduces relation prediction cost to near-linear without substantial recall loss; (iii) explicit multi-step `think` prompts produce structurally richer reasoning traces for LLM-as-a-Judge evaluation. Future work includes fine-tuning a small model on the UN-RES training set, exploring cross-document relation links, and learning a calibrated cosine threshold via cross-validation.

Limitations

Tag recall. With 141 tags and only k=20 candidates pre-filtered by embedding similarity, rare thematic tags may be systematically missed. An ensemble of multiple embedding queries or a hierarchical tag-grouping strategy could improve recall.

Relation and reasoning quality. The cosine similarity threshold θ=0.30 was set empirically without cross-validation, which may affect the precision–recall trade-off. Additionally, while prompts enforce structured reasoning, the quality of generated `think` chains remains uneven. Future work could incorporate learned thresholds or reranking models, along with improved reasoning elicitation techniques.

Reproducibility. Our pipeline depends on INT8 quantisation via `bitsandbytes`, which introduces non-deterministic rounding across hardware. As a result, outputs may vary slightly across GPU architectures.

Acknowledgments

We thank the UZH ArgMining 2026 shared task organisers for providing the dataset and evaluation infrastructure. We also thank Kaggle for providing free GPU compute resources that made this work possible.

References

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Yingjia Gao and others. 2026. Reconstructing the reasoning in UN resolutions: A shared task on argument mining. In *Proceedings of the ArgMining Workshop at ACL 2026*. Shared task description paper.

Yingqiang Gao, Fabian Winiger, Patrick Montjourides, Anastassia Shaitarova, Nianlong Gu, Simon Peng-Keller, and Gerold Schneider. 2025. SpiritRAG: A Q&A System for Religion and Spirituality in the United Nations Archive. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 26–41.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

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

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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 (EMNLP)*, pages 3982–3992. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Appendix

A Hyperparameters

Table A lists the key hyperparameters used in our pipeline.

Component	Parameter	Value
LLM	Model	Qwen2.5-7B
	Quantisation	INT8
	Max new tokens	1,536
	Temperature	0.05
	Top-p	0.9
	Repetition penalty	1.1
	JSON retries	3
Embeddings	Model	mpnet-base-v2
	Dimensions	768
	Batch size	256
Subtask 1	RAG top-k	5
	Tag candidates	20
Subtask 2	Window size w	8
	Cosine threshold θ	0.30

Table A: Key hyperparameters.