TypeCoT at UZH Shared Task 2026: Reconstructing Argumentative Structure in UN Resolutions using Type-Informed Chain-of-Thought

Chandan Kumar R S[1] **Vinay Babu Ulli**[2] **Jyoti Kumari**[3] **Vaibhav Singh**[3]
[1]School of Engineering, Mysore University, Karnataka, India
[2]Oogwai Analytics, Karnataka, India
[3]Department of Linguistics, Banaras Hindu University, Uttar Pradesh, India
chandankumarrs683@gmail.com ullivinaybabu@gmail.com
{jyoti, vaibhav.singh}@bhu.ac.in

Abstract

United Nations and UNESCO resolutions encode complex collective reasoning through highly structured preambles and operative clauses. Reconstructing this implicit argumentative structure is a challenging natural language processing task. This paper describes our submission to the UZH Shared Task at the ArgMining Workshop 2026. Adhering to the strict constraint of using open-weight models with $\leq$ 8B parameters, we propose a highly efficient, modular pipeline built entirely upon the Qwen-2.5-7B-Instruct architecture. To address Subtask 1, we decouple the problem: employing a 4-bit quantized LoRA adapter via the Unsloth framework for paragraph type classification, alongside a dimension-chunked zero-shot approach to assign multi-label tags from a complex 141-class taxonomy. For Subtask 2, we introduce a novel Type-Informed Chain-of-Thought (CoT) methodology that leverages predicted structural metadata as formal constraints to extract argumentative links. To overcome the inherent context window limitations of sub-8B models on extended documents, we implement a multi-pass context recovery pipeline. Our system successfully processes the entire test set of 2,959 paragraphs, ultimately securing an overall Final Rank of 8th in the shared task (ranking 6th on the F1 leaderboard and 8th on the LLM-Judge leaderboard).

1 Introduction

United Nations resolutions are foundational texts in international law and policy. They encode collective reasoning through carefully structured preambular clauses (which provide historical context and justification) and operative clauses (which contain explicit directives and decisions). Extracting these rhetorical frameworks is a specialized sub-field of Argument Mining (Peldszus and Stede, 2013; Lippi and Torroni, 2016). The UZH Shared Task at the ArgMining Workshop 2026 (co-located with ACL 2026) aims to evaluate how well computational systems can reconstruct this implicit reasoning structure from complex legal texts.

The shared task evaluates paragraph-level argumentative structure extraction through two distinct and sequential subtasks:

- **Subtask 1: Argumentative Paragraph Classification.** For each paragraph, systems must predict (a) whether it is *preambular* or *operative*, and (b) assign a subset of 141 predefined education-related tags as a multi-label classification problem.

- **Subtask 2: Argumentative Relation Prediction.** Given a paragraph, systems must predict which other paragraphs it is logically related to (identifying them by their paragraph indices), and label each link with one or more argumentative relation types: *contradictive*, *supporting*, *complemental*, or *modifying*.

A central constraint of this shared task is the restriction to open-weight language models with a maximum of 8 billion parameters. This deliberate limitation prevents reliance on massive proprietary models and forces innovation in prompt engineering, parameter-efficient fine-tuning, and long-context management.

To navigate these constraints, we propose a pipeline built entirely on the Qwen-2.5-7B-Instruct architecture (Qwen Team, 2024). Our system tackles the two official subtasks by chaining structured inferences. Rather than treating relation extraction as an isolated task, the predicted structural metadata from Subtask 1 acts as a constrained prior for relation prediction in Subtask 2. This design builds on methodologies established for parsing persuasive and argumentative discourse (Stab and Gurevych, 2017), ensuring that the language model's inferences are

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 140–144

grounded in the strict logical boundaries inherent to UN documentation. To ensure full reproducibility and support future research, the source code for our multi-stage pipeline is openly available on GitHub at https://github.com/ChandanKumar683/ArgMining. Additionally, all intermediate reasoning predictions and generated datasets are publicly hosted on Hugging Face at https://huggingface.co/collections/Chandan683/argmining.

2 Dataset

Dataset The training data consisted of 2,695 UN Human Rights Council (UN-RES) resolutions (Gao et al., 2025). The test set, however, introduced a deliberate domain and temporal shift, comprising 45 documents (89 JSON files) from UNESCO International Conference on Education (ICE) recommendations spanning 1934 to 2008.

3 System Architecture

To adhere to the 8B parameter limit while managing complex reasoning, we engineered a sequential pipeline built entirely on the Qwen-2.5-7B-Instruct architecture (illustrated in Figure 1).

3.1 Subtask 1: Argumentative Paragraph Classification

Because of the cognitive load required to perform both binary text classification and 141-class multi-label tagging simultaneously, we decoupled Subtask 1 into two independent sub-routines.

Part (a): Paragraph Type Classification For paragraph type classification (*preambular* vs. *operative*), we fine-tuned a Qwen-2.5-7B-Instruct base model using Low-Rank Adaptation (LoRA) (Hu et al., 2022). To maximize computational efficiency, we applied 4-bit quantization via the Unsloth framework, mirroring the QLoRA methodology (Dettmers et al., 2023). The model performs both classification and Chain-of-Thought (CoT) reasoning (Wei et al., 2022) in a single inference pass. The system prompt injects domain expertise, instructing the model to identify *preambular* paragraphs by linguistic markers (e.g., "Recalling...") and *operative* paragraphs by directives (e.g., "Requests..."). The model generates its reasoning inside `<think>...</think>` tags prior to outputting the final label, ensuring interpretable predictions.

Part (b): Multi-Label Argumentative Tagging The multi-label tag set is highly complex, comprising 141 unique codes distributed across 15 distinct dimensions (summarized in Table 1). To reduce hallucination in the 7B model and leverage LLMs as zero-shot reasoners (Kojima et al., 2022), we utilized an approach where tags are assigned **one dimension at a time**. For each of the 15 dimensions, the model receives the relevant tag codes, descriptions, and an explicit "NA" option. This resulted in 15 localized API calls per paragraph (totaling 44,385 inference calls). Valid codes were extracted via parsing, filtering out "NA". Of the 2,959 paragraphs, 2,949 successfully received tags, with the remaining 10 correctly retaining empty lists as transitional fragments.

Dimension	Example Tags	Count
Education level	ISC_1, ISC_23	15
Education orientation	O_G, O_VET	3
Learning modality	M_FORM, M_LL	4
Teachers	T_INI, T_RECR	9
Infra. & resources	INFRA_ICT	8
Curriculum	CUR_DVPMT	7
Pedagogy & assess.	PEDAG_METHO	6
Subject domain	F_MATH, F_SCI	10
Cross-cutting themes	CCUT_DIGIT	11
Policy theme	POL_EQUIT	15
System monitoring	SYST_STAT	7
Legal frameworks	LAW_CONSTI	7
Stakeholder focus	ACT_GOV	8
Learner population	POP_CHILD	12
Ownership/Provision	OWN_PUB	5

Table 1: Overview of the 15 dimensions comprising the 141 unique argumentative tags used in Subtask 1b.

3.2 Subtask 2: Argumentative Relation Prediction

To predict related paragraph indices and their corresponding relation types, we employed a base Qwen-2.5-7B-Instruct model utilizing a novel **Type-Informed Chain-of-Thought** approach. Furthermore, true relations in this corpus are strictly backward-pointing.

Source → Target	Frequency	Dominant Relation
Operative → Operative	61.3%	Complemental (78%)
Preambular → Preambular	22.7%	Complemental (67%)
Operative → Preambular	16.0%	Supporting (89%)
Preambular → Operative	**0.0%**	**Never occurs**

Table 2: Empirical correlation between paragraph types and relations based on development data.

We implemented a two-step reasoning protocol to extract these links:

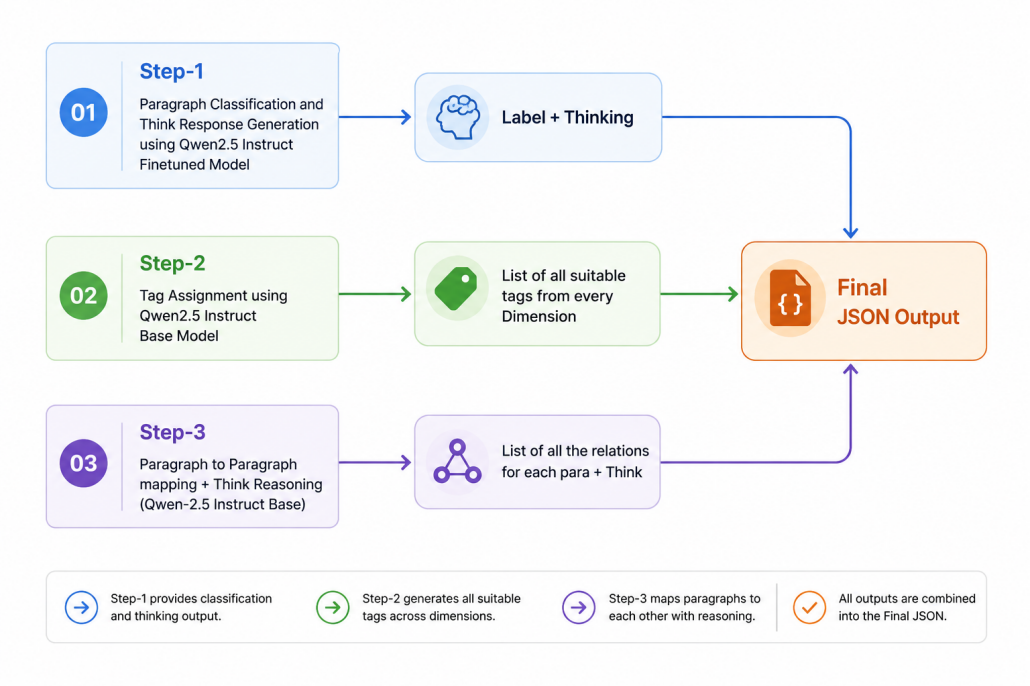

Figure 1: Overview of our three-step system architecture. Step 1 and Step 2 tackle Subtask 1 (Classification and Multi-Label Tagging) using decoupled models, while Step 3 tackles Subtask 2 (Relation Prediction). All outputs are merged into a unified JSON format.

1. **Type-Annotated Summarization:** The model receives the paragraph text alongside its predicted type from Subtask 1a. It produces a structured JSON summary (main claim, key topics, rhetorical verbs, logical dependencies on prior indices).

2. **Type-Constrained Prediction:** The model receives the summaries and the empirical rules from Table 2. It executes a step-by-step reasoning protocol to identify logical candidate indices, verify genuine argumentative links, and classify the relation type (*contradictive, supporting, complemental, modifying*) while explicitly validating against type constraints (e.g., automatically rejecting *preambular* $\rightarrow$ *operative* links).

Post-processing filters were applied to remove self-references, forward references, and out-of-bounds indices, normalizing all relation outputs to lowercase.

3.3 Implementation and Infrastructure

To ensure reproducibility and efficiency under the 8B parameter constraint, our pipeline was implemented using Python 3.10+ and relied on several specialized frameworks. For Subtask 1, local model fine-tuning and base inference were accelerated using the Unsloth framework, which enabled highly efficient 4-bit quantized LoRA training on local GPUs. Data processing and model management were handled via the `transformers`, `datasets`, and `pandas` libraries.

For Subtask 2, managing the extended context window required high-throughput inference. We accessed the base Qwen-2.5-7B-Instruct model via the OpenRouter API (`openrouter.ai/api/v1`) using the `openai` Python client. All intermediate datasets, including the stage-by-stage generated labels and Chain-of-Thought reasoning traces, were stored and hosted using Hugging Face Datasets.

4 Engineering Challenges: Context Limit Recovery

The 32,768-token context window of Qwen-2.5-7B proved to be the primary engineering bottleneck. Documents with over 45 paragraphs frequently exceeded this limit when provided with both text and summarization contexts, resulting in truncated JSON outputs. This aligns with known limitations where models fail to retrieve or process information effectively over extended contexts (Liu et al., 2024).

To achieve 100% coverage, we implemented a three-step **Multi-Pass Recovery Pipeline**:

1. **Main Inference:** Standard Type-informed CoT (2 API calls per document).

2. **Truncated Response Recovery:** Regex-based extraction of complete paragraph objects from exhausted responses, recovering an average of 85% of paragraphs.

3. **Chunked Retry:** Missing paragraph indices were re-processed in batches of 15. The prompt was modified to include only a condensed, one-line summary of previously processed paragraphs, ensuring later chunks could still map relations to earlier document sections without exceeding context bounds.

5 Final Evaluation Results

The TypeCoT system was formally evaluated on the hidden test set by the UZH Shared Task organizers. The evaluation framework utilized two primary tracks: an F1-based metric for strict classification/extraction accuracy, and an LLM-as-a-Judge metric designed to assess the nuanced reasoning and quality of the extracted argumentative relations.

As shown in Table 3, our system achieved a rank of 6th on the F1 Leaderboard and 8th on the LLM-Judge Leaderboard, culminating in an overall Final Rank of 8th among all participating teams.

Metric / Track	**TypeCoT Rank**
F1 Leaderboard	6
LLM-Judge Leaderboard	8
Overall Final Rank	**8**

Table 3: Official shared task evaluation results for the TypeCoT system.

These results successfully validate the effectiveness of our highly constrained ($\leq$ 8B parameters) pipeline. Our stronger comparative performance on the F1 metric (Rank 6) highlights the robustness of our dimension-chunked tagging and Type-Informed CoT extraction methodologies. However, while dimension-chunking mitigated hallucination in Subtask 1, qualitative observations suggest that tagging performance is heavily dependent on the dimension. Dense dimensions with distinct semantic boundaries (e.g., *Subject domain*) are more reliably predicted in a zero-shot setting than highly granular or overlapping dimensions (e.g., *Cross-cutting themes*).

The slight drop in the LLM-Judge ranking (Rank 8) aligns directly with our internal observations regarding the 7B model's behavior in Subtask 2. Specifically, the model exhibits a tendency toward relational over-connection (linking paragraphs to too many prior indices) and the under-prediction of minority classes (such as *modifying* and *contradictive*). This suggests that while our type-informed constraints successfully enforced structural compliance, they were likely too rigid, unintentionally sacrificing minority class recall. While an F1 metric may partially tolerate dense graph extraction if the true positives are captured, a nuanced LLM judge naturally penalizes this lack of relational precision.

6 Conclusion

We presented a comprehensive two-subtask pipeline for the UZH ArgMining 2026 Shared Task that strictly adheres to the $\leq$ 8B parameter constraint. By leveraging parameter-efficient fine-tuning for paragraph classification, dimension-chunked zero-shot multi-label tagging, and a Type-Informed Chain-of-Thought approach backed by a context-recovery pipeline, we demonstrated a robust method for extracting directed, labeled relational graphs from complex legal texts using highly efficient open-weight models.

7 Limitations and Future Work

1. **Domain Shift Exploration:** The LoRA adapter (Subtask 1a) was trained on UN HRC resolutions but tested on older UNESCO ICE recommendations (1930s-1940s). Such temporal and stylistic disparities introduce a well-documented domain shift penalty (Ramponi and Plank, 2020). Future work must explore this shift more deeply, analyzing how the historical evolution of rhetorical markers degrades heuristic-based retrieval.

2. **Zero-Shot Tagging Recall:** While dimension-by-dimension prompting reduced hallucination in Subtask 1b, zero-shot assignment typically achieves limited recall. Fine-tuning specifically for multi-label classification is recommended, alongside a rigorous, per-dimension performance analysis to identify where zero-shot reasoning fails.

3. **Ablation of Heuristic Constraints:** Our approach heavily relies on prompt design and explicit structural heuristics. A critical next step is to conduct controlled ablation studies comparing our Type-Informed CoT against a simpler, unconstrained baseline to accurately quantify the robustness and isolated contribution of these rules.

4. **Model Constraints:** The 8B parameter limit restricted the ability to natively parse entire long documents with deep logical reasoning. While Qwen-2.5-7B provided a strong baseline, its context window necessitated aggressive summarization via our multi-pass pipeline.

5. **Class Imbalance Mitigation:** The prompt-based structural guidance for Subtask 2 systematically suppressed minority relation types, indicating our constraints were potentially unbalanced and too strong. Implementing targeted few-shot examples (Brown et al., 2020) or a post-hoc reclassification pass focusing on constraining linguistic markers could recover *modifying* and *contradictive* relations.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th international conference on computational linguistics*, pages 6838–6855.

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, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.