HybridArguer at UZH Shared Task 2026: Argument Structure Modeling in Bilingual UN Resolutions with Retrieval-Augmented and Iterative LLM Reasoning

Siddharth Bhargava
Fondazione Bruno Kessler / Trento Italy
Universidade da Coruña / A Coruña Spain
sbhargava@fbk.eu

Abstract

Extracting argument structures from legal-political discourse reveals how policies and actions are proposed, debated, and formalized, but remains challenging due to the complexity of long-form, structured text. This work proposes a modular, retrieval-augmented system for traceable and structured argument mining in long, bilingual United Nations resolutions.

This paper describes our system submission to the UZH Shared Task 2026, focusing on practical design choices for argument structure modeling under task and model constraints. Our system employs a parameter-efficient ($\leq$8B) open-source model, Qwen3:8B in *thinking* mode, to perform paragraph classification, multi-label tag assignment, and multi-label relation prediction through a modular, retrieval-augmented pipeline. Our approach integrates retrieval augmentation with self-consistency and self-refinement to address challenges related to scale, no supervision, structured reasoning, and output stability. Experiments on the dataset show that retrieval augmentation improves tag assignment precision, while iterative prompting enhances consistency in output prediction across documents of varying length. The code has been made publicly available [1].

1 Introduction

Argument mining (AM) is a key component of legal-political discourse analysis for downstream tasks such as policy decision-making and conflict resolution (Heinisch et al., 2024; Plenz et al., 2024). It presents a challenging setting, being grounded in formal argumentation theory and structured schemata while remaining highly sensitive to discourse cues and contextual relationships, particularly in **long-form discourse** such as the legal-political documents (Mao et al., 2024; Feger et al., 2025). This raises a fundamental question of how computational models can effectively apply theoretical concepts to long-form, real-world data, rather than relying on surface-level statistical patterns that lead to limited task understanding (Feger et al., 2025).

Prior work on argumentation in long-form discourse has explored approaches based on textual features (e.g., TF-IDF (Huwaidah et al., 2021)), embedding-based methods (Carlebach et al., 2020), document chunking via rolling windows (Quijano Sánchez and Cantador Gutiérrez, 2020), and graph-based formulations that model semantic relationships across text units (Dore et al., 2025). However, these approaches often face a trade-off between constructing rich, reliable structured representations and the level of human effort required for modeling and supervision, limiting their scalability and application. Additionally, scaling such methods to long-form discourse is often computationally expensive (Martinelli, 2025).

Large language models (LLMs) have recently emerged as a promising alternative, offering strong contextual understanding with minimal task-specific supervision (Kojima et al., 2022). Earlier LLMs primarily rely on implicit pattern recognition and often struggle to consistently apply structured reasoning, limiting their reliability in argument mining tasks (Feger et al., 2025). The emergence of *reasoning* LLMs (Wei et al., 2022) in explicit thinking mode addresses this limitation by exposing intermediate reasoning traces explicitly, enabling closer inspection of how models identify discourse markers and infer relationships between argumentative units. However, a systematic analysis remains limited.

Along this line of research, the UZH Shared Task on Reconstructing the Reasoning in United Nations Resolutions, organized by the University of Zurich (UZH) under the 13th Workshop on Argument Mining (2026), investigates how computational systems can effectively recover the underly-

[1] https://github.com/The-obsrvr/ArgStructurePredictionInUNResol

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 151–159

ing argument structure of long, bilingual UN resolution documents. Specifically, systems must analyze each paragraph to: (i) classify it as *preambular* or *operative*, (ii) assign descriptive tags, and (iii) predict its relationships—along with their types—with other paragraphs. The task is constrained to lightweight open-source models ($\leq$8B parameters), introducing challenges such as long hierarchical documents, multi-label annotation, and structured relation prediction under limited supervision.

In this work, we propose a modular pipeline that decomposes these subtasks and enables efficient and structured argument modeling. Given the large tag space (141 predefined tags) and the substantial number of possible paragraph matches (on average 700 possible matches per document), we adopt a conservative, retrieval-augmented prompting strategy that first narrows down tag and source paragraph candidates for each target paragraph, followed by a zero-shot LLM-based selection of final tags and relations (Efeoglu and Paschke, 2025). We investigate how the reduced search space on tags and paragraphs facilitates LLM reasoning while minimizing the computational demand.

To align with the requirements of the Shared Task, we explicitly capture the model's reasoning traces for classification and tagging. Moreover, in the absence of supervised data and to improve the consistency and structural validity of these outputs, we adopt self-consistency (Wang et al., 2022) and self-refinement (Madaan et al., 2023) through iterative prompting to support LLM reasoning. Finally, we leverage bilingual signals by combining the original French text with the provided English translations to construct richer textual embeddings (Wang et al., 2024), which enhance the effectiveness of the candidate selection modules (Ranaldi et al., 2026).

The main contribution of this work is a modular, retrieval-augmented end-to-end framework for traceable, structured argument mining in long bilingual UN resolutions, covering paragraph-level classification, multi-label tag assignment, relation prediction, and multi-label relation-type classification.

2 Data Description

The dataset used in this work is provided as part of the Shared Task and taken from the UN-RES dataset (Gao et al., 2025). It contains paragraph-level argumentative structures. The training set comprises 2,695 bilingual (French–English) documents, while the held-out test set includes approximately 45 resolutions distributed across 90 JSON files following a strict, predefined schema.

The training data includes annotations for paragraph classification; however, reliable ground truth labels for tag assignment and relation prediction are not available. This limitation motivates the use of consistency-based strategies—through iterative prompting—and the reasoning capabilities of modern large language models to perform downstream tasks in zero-shot or weakly supervised settings.

3 Methodology

The system comprises four main modeling stages: (1) a reasoning LLM classifies paragraphs collectively as *preambular* or *operative*; (2) embedding-based similarity retrieves tag candidates for each paragraph; (3) candidate source paragraphs are similarly selected under a chronological constraint; and (4) the LLM processes each target paragraph individually to assign tags from its candidate pool and predict no, one or more predefined relation types with candidate source paragraphs (see Figure 1).

1. Document-level LLM Prompting. For each document, paragraphs in their French form are truncated to the first 120 characters and concatenated into a paragraph list. We employ a reasoning LLM in **thinking** mode, prompted in a zero-shot setting to classify each paragraph as *preambular* or *operative*, following their definitions in UN Editorial Manual (United Nations, n.d.). Intermediate reasoning traces are extracted using the `</think>` marker. To improve robustness, self-consistency is applied by running two prompts with low temperatures (0.10 and 0.15) and merging outputs via majority voting. Disagreements were resolved through a lightweight tie-breaking strategy. Because UN resolutions consistently present preambular paragraphs before operative ones, we maintained an *operative* flag initialized to `false`. The flag was activated only when both prompt generations classified a paragraph as operative. During disagreements, the system defaulted to the *preambular* label while the operative flag remained inactive, and to *operative* otherwise. This heuristic reduced the need for a third prompt generation while remaining effective for this relatively trivial classification task. Additionally, in cases of failed generation after multiple attempts, a fallback heuristic assigns labels based on semantic discourse markers, with

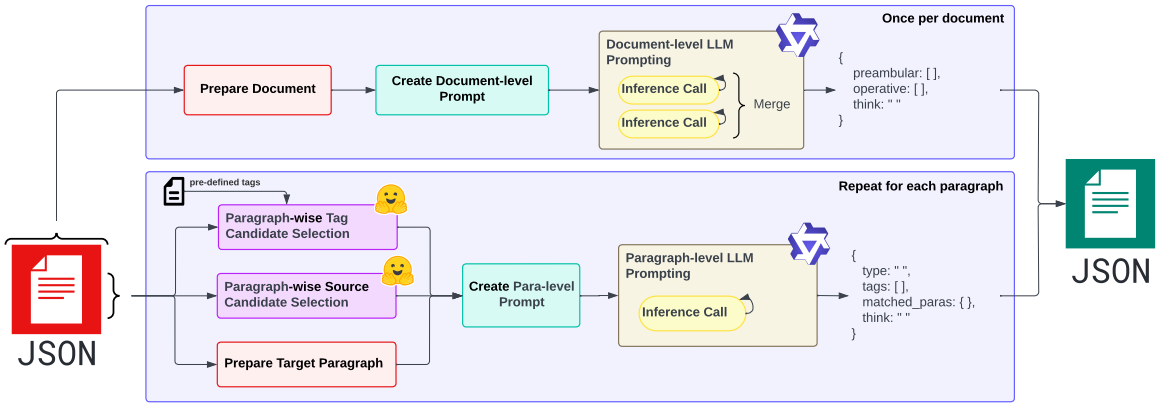

Figure 1: **System Architecture**: The figure illustrates the four-stage end-to-end pipeline that processes a JSON document to produce a structured output. Purple modules correspond to candidate selection stages (Stages 2 and 3), where a multilingual embedding model retrieves top tag candidates and source paragraph candidates for each target paragraph. Yellow modules represent LLM prompting stages (Stages 1 and 4), where the model performs reasoning-based predictions for classification, multi-label tagging and multi-label relation prediction.

corresponding reasoning recorded.

2. Paragraph-wise Tag Candidate Selection. We use *multilingual-e5-large* (Wang et al., 2024), a popular multilingual text embedding model on Hugging Face, to generate embeddings for each tag based on its dimensional and categorical descriptions. For each paragraph, query embeddings are computed as the normalized average of its English and French representations. K-Nearest Neighbors (KNN) with cosine similarity is used to retrieve top 20 tag candidates, further filtered down using a similarity threshold (≥ 0.45) returning top 5 candidates. We use this relatively conservative setup to prioritize recall over precision.

3. Paragraph-wise Source Candidate Selection. Similarly, embeddings for all paragraphs are constructed by averaging their English and French representations. The chronological constraint defined in the Shared Task is enforced, requiring source paragraphs to appear after the target paragraph, thereby ensuring directional consistency in relation prediction. KNN with cosine similarity (≥ 0.45) is used to retrieve top candidate source paragraphs for each target paragraph. To ensure local dependencies that likely may be related to the target, up to three immediately following paragraphs are additionally included if not selected via similarity.

4. Paragraph-level LLM Prompting. The reasoning LLM is again employed and prompted per target paragraph. The model assigns tags from the candidate pool (Stage 2) in a multi-label classification setting, with brief justifications. Within the same prompt, it predicts if a relation(s) exists with each candidate source paragraph (Stage 3) and then classifies them with one or more of four predefined types: *supporting, contradicting, modifying,* and *complemental*.

The model is encouraged to rely on discourse cues when predicting relations and to produce confidence scores, for improved reasoning. Up to three generation attempts are allowed, with simple corrective feedback applied in case of invalid outputs. Self-consistency is not used at this stage due to the significantly high computational cost associated with paragraph-level processing.

Outputs from Stages 1 and 4 are integrated into the final structured format required by the Shared Task. Additional implementation details on the prompts are provided in the Appendix A.

4 Experimental Results

All experiments were conducted on a single 48GB NVIDIA A40 GPU. To our knowledge, only a limited number of open-source, lightweight LLMs provide an explicit reasoning (exclusive "thinking" mode) capability within the $\leq$8B parameter range. For this study, we evaluated the following models: Qwen3:4B-Thinking-2507 (Yang et al., 2025), Qwen3:8B (Yang et al., 2025), operated in explicit thinking mode; and DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), which exhibits implicit reasoning behavior. The models have been evaluated without introducing iterative prompting, except for Qwen3:8B which presents both scenarios.

Model	Para. Cls.			Tag Assign.			Rel. Pred.		
	Acc.	Stab.	Reason	Acc.	Stab.	Reason	Acc.	Stab.	Reason
Qwen3:4B-Think	**0.90**	0.35	0.25	0.40	0.15	0.45	0.45	0.30	0.55
Qwen3:8B (no IP)	0.70	0.30	0.40	0.60	0.45	**0.65**	0.55	0.55	**0.70**
Qwen3:8B (IP)	0.78	**0.55**	**0.60**	**0.70**	**0.65**	0.62	**0.60**	**0.72**	0.55

Table 1: Performance of models across three tasks, as evaluated by an LLM-based judge (GPT-5). All scores are reported on a 0–100 scale and measure accuracy (Acc.), output stability (Stab.), and reasoning quality (Reason) across the three main tasks: paragraph classification (Para Cls.), multi-tag assignment (Tag Assign.) and multi-relation prediction (Rel. Pred.). IP denotes iterative prompting.

Due to the absence of ground truth annotations for the multi-label tag assignment and relation prediction tasks, a larger LLM, GPT5 (Singh et al., 2025), was employed as an LLM-as-a-judge to evaluate a subset of 5 documents (around 100 paragraphs) sampled from the held-out test set. In addition, manual self-evaluation was conducted to perform error analysis of the predictions (Appendix B). The evaluation focuses on the observed performance across three tasks: (i) paragraph classification, (ii) multi-label tag assignment, and (iii) multi-label relation prediction and classification. This combined evaluation setup was used to determine the final system configuration, which was then applied to the full held-out test set.

The official evaluation framework of the Shared Task adopts both a F1 score and an LLM-as-a-judge metric to assess system performance in terms of prediction accuracy and reasoning quality respectively.

4.1 Exploratory Model Evaluation

Table 1 summarizes model performance across the three tasks, as evaluated by an LLM-based judge (see Appendix A.3 for the evaluation prompt). We consider three evaluation metrics: *accuracy* (Acc.), *stability* (Stab.), and *reasoning quality* (Reason). Accuracy reflects how correctly a model performs each task according to the LLM judge. Stability measures the frequency with which fallback logic is triggered, indicating failures to produce valid structured outputs. Reasoning quality evaluates the coherence and usefulness of the reasoning traces generated during successful inference.

DeepSeek-R1-Distill-Qwen-7B failed to reliably execute the reasoning step under structured generation, resulting in consistently poor outputs. Post-hoc analysis revealed that this behavior was primarily caused by a prompting template originally designed for the Qwen3 architecture, which proved incompatible with the model's reasoning format. Consequently, this configuration was omitted from the reported results.

For the paragraph classification task, we observed higher accuracy with the smaller Qwen3:4B model. Manually inspecting the Qwen3:4B it was better at handling a logical error made by us, where bullet point markers were defined incorrectly in the prompt leading to mis-classification of operative paragraphs. The larger Qwen3:8B variant appeared more sensitive to this distribution shift, leading to reduced classification performance. In terms of stability, the Qwen3:8B iterative prompting (IP) configuration reduced reliance on fallback logic, whereas the non-IP variants struggled to consistently generate valid structured outputs. The higher rate of successful generations also improved the quality of the resulting reasoning traces.

For the tag assignment task, the IP configuration again demonstrated the strongest performance. Manual inspection showed improved task alignment, with the model successfully identifying multiple relevant tags per paragraph. However, the models also exhibited a tendency to overuse default categories, particularly the "N/A" tag within several dimensions. Despite this, the generated reasoning traces generally provided coherent and interpretable explanations for the assigned tags, suggesting that the models captured the task semantics reasonably well. Self-consistency was not applied at this stage due to computational cost considerations; however, self-refinement remained active and contributed to improved generation stability.

The relation prediction task exhibited substantially greater variability than the previous two tasks. Although the IP configuration again achieved the highest stability, the models frequently struggled to assign diverse relation labels and often defaulted to predicting the "complemental" relation type. While

the LLM did not conform all candidate source paragraphs into a relation, the observed tendency to over-predict "complemental" relations suggests a potential bias in the reasoning process. Due to the absence of reliable ground-truth label distributions, it remains difficult to determine whether this behavior reflects genuine task alignment or an LLM prediction error.

Overall, the evaluated systems demonstrated moderate structural understanding across the three tasks. Iterative prompting and self-consistency improved both output stability and reasoning quality for Qwen3:8B, although the trade-off between computational cost and performance gains warrants further investigation. Among the tasks, paragraph classification proved the most challenging in terms of producing stable structured outputs. Nevertheless, its deterministic logical structure allowed fallback mechanisms to recover reasonably accurate predictions. Only Qwen3:8B (IP) achieved partial success on some documents without requiring fallback logic (see Appendix B for manual inspection summary).

5 Discussion

Qwen3:8B with iterative prompting achieved the strongest overall performance among the evaluated configurations. Its explicit reasoning mode improved both output stability and task accuracy, particularly for structured prediction tasks such as tag assignment and relation prediction. Consequently, it was selected as the primary reasoning model within our pipeline (Figure 1) to generate our final submissions.

In the Shared Task evaluation, our system (*HybridArguer*) ranked third overall (4th by F1 and 3rd by LLM-as-a-judge evaluation). According to the organizers' silver evaluation, the system achieved a type macro F1 score of 0.891 and a relation label F1 score of 0.389, while obtaining a notably high pair recall of 0.713. The discrepancy between pair recall and relation type classification performance suggests that, although the system reliably identified relevant paragraph pairs, it was less effective at assigning precise relation labels. Manual inspection indicated that the LLM frequently over-generalized toward dominant relation categories—particularly "complemental"—or misinterpreted fine-grained relation semantics during label assignment.

Additionally, despite explicit prompting instructions permitting multi-label relation assignment, the evaluated LLMs consistently produced only a single relation label per paragraph pair. Interestingly, this behavior did not negatively affect the tag assignment task, where multi-tag predictions remained comparatively reliable. These observations suggest that relation prediction is relatively a more difficult challenge than paragraph-level tagging.

Overall, the results demonstrate the effectiveness of combining retrieval augmentation with iterative LLM reasoning, while highlighing to the need for better supervision and prediction alignment.

6 Conclusion

We demonstrate how long, bilingual documents can be efficiently analyzed to produce traceable and structured argument representations. Our proposed system enables modular processing, facilitating improved control, interpretability, and scalability across diverse document settings.

This work serves as a preliminary step toward robust structured argument mining in long-form, legal-political discourse. By leveraging retrieval augmentation and iterative prompting, the proposed approach addresses key challenges related to scale, complex reasoning, and output stability.

Future work should focus on incorporating supervised training to improve label accuracy and alignment with ground truth. Additionally, integrating knowledge graph-based approaches for modeling document structure presents a promising direction for enhancing reasoning consistency and relation prediction (Dore et al., 2025; Muniraja and Satapathy, 2026). Finally, adopting more extensive evaluation strategies that combine human judgment with automated metrics may provide a more reliable assessment of structured reasoning performance in complex document settings.

Limitations

First, this work employs relatively simple and conservative retrieval conditions, prioritizing recall over precision, such as setting similarity threshold to ≥ 0.45 without exploring alternative options, and setting maximum final candidates to top 5 upon which the LLM then reasons and selects the final values. While this design choice ensures qualitative candidate selection, it may limit identifying sensitive or noisier candidates in downstream predictions, especially given the multi-label conditions. Future work should explore more refined

retrieval strategies to better balance precision and recall across tasks.

Second, the evaluation is limited by the absence of comprehensive ground truth for all subtasks, relying instead on a small-scale evaluation process, primarily done by the author and a larger LLM model. A more rigorous and large-scale evaluation framework is necessary to reliably assess the system's accuracy, consistency, and robustness across diverse settings.

Third, while retrieval augmentation help reduce computational cost, there is still quite a computational demand required to analyze long documents. For instance, a document containing 70 paragraphs would require making 71 inference calls to the model (considering no iterative prompting was employed). Thus, future work must focus on further reducing the computational cost, for instance using clustering techniques to analyze similar target paragraphs collectively rather than through an individual process. Additionally when including iterative prompting, it introduces additional computational cost and latency. The tradeoff between the computational cost and the performance gain remains to be studied.

Finally, the proposed approach has not been extensively compared against strong state-of-the-art baselines. Future work should include systematic benchmarking against state-of-the-art models and methods to better compare the system in terms of performance and computational efficiency.

Acknowledgements

This research work has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

References

Mark Carlebach, Ria Cheruvu, Brandon Walker, Cesar Ilharco Magalhaes, and Sylvain Jaume. 2020. News Aggregation with Diverse Viewpoint Identification Using Neural Embeddings and Semantic Understanding Models. In *Proceedings of the 7th Workshop on Argument Mining*, pages 59–66, Online. Association for Computational Linguistics.

Deborah Dore, Stefano Faralli, and Serena Villata. 2025. Leveraging Graph Structural Knowledge to Improve Argument Relation Prediction in Political Debates. In *Proceedings of the 12th Argument Mining Workshop*, pages 74–86, Vienna, Austria. Association for Computational Linguistics.

Sefika Efeoglu and Adrian Paschke. 2025. Fine-Tuning Large Language Models for Relation Extraction within a Retrieval-Augmented Generation Framework. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 1–7, Vienna, Austria. Association for Computational Linguistics.

Marc Feger, Katarina Boland, and Stefan Dietze. 2025. Limited Generalizability in Argument Mining: State-Of-The-Art Models Learn Datasets, Not Arguments. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23900–23915, Vienna, Austria. Association for Computational Linguistics.

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, Suzhou, China. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Philipp Heinisch, Lorik Dumani, Philipp Cimiano, and Ralf Schenkel. 2024. "Tell me who you are and I tell you how you argue": Predicting Stances and Arguments for Stakeholder Groups. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1968–1982, Mexico City, Mexico. Association for Computational Linguistics.

Amalia Huwaidah, Adiwijaya, and Said Al Faraby. 2021. Argument Identification in Indonesian Tweets on the Issue of Moving the Indonesian Capital. *Procedia Computer Science*, 179:407–415.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 22199–22213, Red Hook, NY, USA. Curran Associates Inc.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon,

Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *Preprint*, arXiv:2303.17651.

Tiezheng Mao, Osamu Yoshie, Jialing Fu, and Weixin Mao. 2024. Seeing both sides: Context-aware heterogeneous graph matching networks for extracting-related arguments. *Neural Computing and Applications*, 36(9):4741–4762.

Giuliano Martinelli. 2025. Extending coreference resolution to long texts: From paragraphs to full books and beyond.

Pasupuleti Muniraja and Shashank Mouli Satapathy. 2026. KGERA: Knowledge graph enhanced reasoning architecture for recommendation systems. *Scientific Reports*.

Moritz Plenz, Philipp Heinisch, Anette Frank, and Philipp Cimiano. 2024. PAKT: Perspectivized Argumentation Knowledge Graph and Tool for Deliberation Analysis. In *Robust Argumentation Machines*, pages 89–107, Cham. Springer Nature Switzerland.

Lara Quijano Sánchez and Iván Cantador Gutiérrez. 2020. Structured argumentation modeling and extraction: Understanding the semantics of parliamentary content.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. 2026. Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Question Answering Task. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 697–716, Rabat, Morocco. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, A. J. Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. OpenAI GPT-5 System Card. *Preprint*, arXiv:2601.03267.

United Nations. n.d. *United Nations Editorial Manual*. United Nations. Available at: https://www.un.org/dgacm/en/content/editorial-manual.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *Preprint*, arXiv:2402.05672.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

A Further Implementation Details

Clear instructions have been provided in the Github repository describing how the different configurations can be implemented.

Additionally below we provide a simplified draft of the prompt template used to instruct the LLM.

A.1 Prompt for Document Level Prompting

The prompt used to instruct the reasoning LLM to perform document-level prompting is mentioned in Figure 2.

A.2 Prompt for Para-level Prompting

The prompt used to instruct the reasoning LLM to perform document-level prompting is mentioned in Figure 3.

A.3 Prompt for System Evaluation

The prompt used to instruct the LLM judge is mentioned in Figure 4.

B Manual Inspection

Beginning with the paragraph classification task, even with the best performing system, the stability was quite low with the LLM often failing in its structured reasoning and relying on the fallback logic. Given that preambular and operative classes have a clear distinct structure, fallback logic worked relatively well resulting in high accuracy.

For the multi-label tag assignment task, there was no empty outputs by our best performing system. However, on manual inspection it felt that the system defaulted to certain tag patterns or relying on the "nan" option which indicates no specific category under a particular dimension. Thus, tagging potentially may be only surface-level and not semantically grounded.

You are an expert in UN resolution analysis written in French.

Task: (1) Identify how preambular and operative paragraphs are distinguished in this document. (2) Use discourse markers and linguistic cues. (3) Apply one consistent rule to classify all paragraphs. Return strict JSON.

Definitions:
Preambular paragraphs (French): Provide context, justification, or background; may begin with "Considérant", "Rappelant", "Reconnaissant", "Notant", "Soulignant"; often end with commas; do not contain actions.
Operative paragraphs (French): Contain actions, recommendations, or directives; may include verbs such as "Décide", "Demande", "Encourage"; may be structured, numbered, and action-oriented.

Important: Think briefly inside `<think></think>`, then answer.
Numbered paragraphs (1., I., II.) are operative. Usually, after the first operative paragraph, all following paragraphs are operative.

Output (strict JSON format):
```
{ "preambular_para": [list of paragraph numbers], "operative_para": [list of paragraph numbers],
"think": "One unified explanation of how you distinguished preambular vs operative" }
```

Rules: Do not omit any paragraph; each must appear exactly once. Each paragraph must be labeled either preambular or operative. The think field must explain the rule used. Output only valid JSON.

Input: `{para_block}`

Figure 2: Prompt for classifying preambular and operative paragraphs in French UN resolutions.

Finally, multi-label relation prediction was the toughest challenge in terms of accuracy. Relation types were not interpreted correctly. Complemental relations were overly used, while contradictive and modifying relations were observed to be incorrect. It may be that the LLM assigned the relations superficially. Confidence scoring did not explicitly seem to help the reasoning either.

Overall, the manual evaluation suggest that the LLM cannot reliably produce coherent argument structures. However, strategies such as retrieval augmentation that reduces document complexity and iterative prompting, that stablises output generation, is definitely a step in the right direction. We note that this manual evaluation is limited to only 5 documents and a larger more intensive review is needed to make more reliable claims.

You are an expert in UN resolution analysis.

Task: For the given target paragraph: (1) Assign relevant tags from the provided candidate tags; (2) Identify which other paragraphs are meaningfully related; (3) Assign one or more relation types for each related paragraph.

Input (Target Paragraph):
```
para_number: {para["para_number"]}
text_fr: {para["para"]}
text_en: {para["para_en"]}
candidate_tags: {tag_block}
relation_candidates: {candidate_para_block}
allowed_paragraph_ids: {relation_candidates}
```

Multiple Tag Classification: Select only from candidate tags. Output only tag codes (e.g., "A1", "B2"). A paragraph may have multiple tags. Include only tags clearly supported by the paragraph content. Explain clearly why the tags have been selected.

Relations (Multi-label): A pair of paragraphs may have multiple relation types. Assign all relation types that are clearly supported.

Relation Decision Process (Very Important): For each candidate paragraph:
Step 1: Check if a meaningful relation exists. If there is no clear semantic connection, do not include the paragraph.
Step 2: If a relation exists, evaluate each relation type independently: supporting (reinforces or justifies), contradictive (opposes or restricts), modifying (refines or adds conditions), complemental (adds related but independent information).
Step 3: Assign a confidence score and include all relation types with confidence 0.5: 0.9–1.0 very strong; 0.7–0.9 strong; 0.5–0.7 moderate; < 0.5 weak (exclude). Confidence must reflect how clearly the relation is supported by the text.

Reasoning Guidelines: Think carefully before selecting. For each relation: identify the idea in the target paragraph, identify the corresponding idea in the candidate paragraph, base the relation only on these aligned parts.

Output (Strict JSON Format):
```
{
"para_number": {para["para_number"]},
"tags": ["tag_code_1", "tag_code_2", ...],
"matched_paras": {"X": [{"type": "relation_type", "confidence": 0.0}]},
"think": "Briefly explain why the selected tags apply"
}
```

Output Rules: Output only valid JSON. Do not include `<think>` tags. The final answer must start with `{` and end with `}`. Do not return empty JSON `{}`.

Constraints: Use only tag codes from candidate tags. Use only paragraph IDs from allowed paragraph IDs. Each relation must include a confidence score. Do not include relations with confidence < 0.5. Do not include the target paragraph itself. Do not include empty relation entries.

Figure 3: Prompt for multi-label tag assignment and relation prediction between paragraphs.

You are a strict LLM-as-a-judge and an expert in argumentation mining for UN resolution analysis. Carefully review and score (0–100) the following system outputs based on output consistency (stability), reasoning quality, and accuracy, using both the final outputs and the provided reasoning traces.

(1) Paragraph Classification (Preambular vs. Operative). Assess how well the system annotates paragraphs. Accuracy (0–100): Are paragraphs correctly classified? Stability (0–100): Has the output relied on fallback logic due to failed structured generation? Reasoning Quality (0–100): How well are the reasoning traces (in the METADATA section) generated?

(2) Multi-label Tag Assignment. Assess how well the system assigns tags to each paragraph. Accuracy (0–100): Are the tags semantically relevant to the paragraph? Stability (0–100): Has the output relied on fallback logic returning an empty list due to failed generation? Reasoning Quality (0–100): How well do the reasoning traces (in the BODY section) explain the tag assignment?

(3) Multi-label Relation Prediction. Assess how well the system predicts relations between paragraphs. Accuracy (0–100): Are the relations (*matched_paras*) semantically relevant and correctly labeled? Stability (0–100): Has the output relied on fallback logic returning an empty dictionary due to failed generation? Note: the last 1–2 paragraphs may not have outgoing relations due to directionality constraints. Reasoning Quality (0–100): How well do the reasoning traces (in the BODY section) explain the relation prediction?

Instructions: Be strict and consistent. Penalize hallucinations, incorrect logic, weak justification, and fallback reliance. Reward precise, consistent, and well-justified reasoning. Use the full 0–100 scale. Return the evaluation clearly and explicitly.

Figure 4: LLM-as-a-judge prompt used for evaluating model outputs across three tasks.