POINTERS at UZH Shared Task 2026: Reasoning Probes for Argumentation Mining in UN Resolutions*

Sohom Sen, Avina Nakarmi, Xun Song, Aritra Dasgupta
Department of Data Science
New Jersey Institute of Technology
{ss4887, an778, xs29, aritra.dasgupta}@njit.edu

Abstract

This paper describes the submission of team **POINTERS** to the UZH ArgMining 2026 Shared Task, which aims to recover the argumentation structure of UN and UNESCO resolutions by labeling paragraph types, assigning specific tags, and predicting relations between paragraphs. We take a generative approach, treating each resolution as a sequence of claim-evidence pairs connected by explicit reasoning strategies. First, each paragraph is classified as *preambular* or *operative* and assigned tags from a 126-code vocabulary, with the model required to quote specific phrases to justify every decision. Second, for each paragraph, we first retrieve semantically related candidates using sentence transformers, then use reasoning strategies as a diagnostic scaffold to label the relation—*supporting*, *complemental*, *contradictive*, or *modifying*—along with a quoted, strategy-grounded rationale. Both steps run locally on **Qwen3-8B-GGUF** (Team, 2025) (NVIDIA RTX 4080, 16 GB VRAM) without any cloud API calls. In the absence of labeled data, we use Claude Sonnet 4.6 only for an internal diagnostic evaluation of the generated reasoning traces. The results show that a sub-8B open-source model can produce evidence-grounded explanations for formal diplomatic text, while relation labeling remains sensitive to the distinction between retrieval and reasoning strategy-based diagnosis.

1 Introduction

UN and UNESCO resolutions follow a recognizable structure: a block of *preambular* paragraphs that recall past agreements and establish context, followed by *operative* paragraphs that issue directives and recommendations (Di Carlo, 2013). On the surface, this structure looks tidy, but the argumentative connections between paragraphs, i.e., why one preambular clause supports a particular operative request, or how two clauses qualify each other, are rarely made explicit. Recovering these connections automatically is the goal of the UZH ArgMining 2026 Shared Task, and it remains a difficult problem (Lippi and Torroni, 2016; Stede et al., 2019). The task asks systems to annotate each paragraph with a type (*preambular* or *operative*), a set of education-dimension tags, and directed argumentative relations to other paragraphs. The task defines two tracks: one focused on producing structured labels directly, and another that additionally requires free-text reasoning trace (*think* field) justifying each decision.

We developed two systems for this task and present here the generative approach, which participates in the reasoning-trace track. Team **POINTERS** chose this track because the reasoning traces make the system's decisions interpretable and provide a richer signal for evaluation. Our starting point is evidential reasoning (Toulmin, 2003; Vaidya and Dasgupta, 2020), which describes how a reader moves from observed evidence to a supported claim via systematic reasoning strategies. We adapt this to the resolution domain: preambular paragraphs play the role of evidence, operative paragraphs are the claims they support, and the four relation types in the task map cleanly onto four reasoning strategies—Causal, Corroboration, Contrastive, and Triangulation. This framing gives the model a principled vocabulary for explaining its decisions rather than labeling in a vacuum.

To keep the system accessible, we run the generation entirely on a local **Qwen3-8B-GGUF** model (Team, 2025) on a single consumer GPU, using Claude Sonnet 4.6 only for evaluation. The system scores 77/100 on the test set under LLM-as-a-Judge evaluation (Zheng et al., 2023), suggesting that structured prompting with explicit reasoning strategies can compensate meaningfully for the limited capacity of a sub-8B model.

*Code available at https://github.com/SenSohom/UZH-Shared-Task-ArgMining-Workshop-2026

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 145–150

2 Tasks and Data

The shared task provides UNESCO International Conference on Public Education (ICPE) resolutions in French, each accompanied by an English translation. Every resolution is stored as a JSON file with a list of numbered paragraphs. Systems must fill three fields for each paragraph.

Paragraph type. Each paragraph is either `preambular` or `operative`. Preambular paragraphs open with contextual keywords: *Recalling, Noting, Convinced, Welcoming*. They explain the rationale behind the resolution. Operative paragraphs open with directive keywords: *Requests, Urges, Decides, Recommends*. They constitute the actual decisions.

Tag assignment. Each paragraph is assigned one or more codes from a controlled vocabulary of **126 education-dimension tags**. Tags cover actors (e.g., `ACT_IO` for international organisations), legal instruments (`LAW_INTER`), and policy themes (`POL_EQUIT` for equity). They capture both the thematic content and the argumentative function of the paragraph.

Argumentative relations. For any pair of paragraphs that are argumentatively connected, the system assigns one or more relation types: **supporting** (one provides the premise that justifies the other), **complemental** (both assert the same claim through different evidence), **contradictive** (one limits or opposes the other), and **modifying** (one adds conditions or exceptions to the other) (Stab and Gurevych, 2014; Habernal and Gurevych, 2017). Relations are stored in a `matched_pars` field keyed by target paragraph number.

Data. The training split contains **2,694 resolution JSON files** spanning ICPE proceedings from 1934 onwards (Gao et al., 2025). The held-out test set has **89 resolutions** with all annotation fields left blank. With no gold labels, we rely on LLM-as-a-Judge (Zheng et al., 2023) for our analysis.

3 Framework and Pipeline

Our starting point is the observation that resolutions are structured arguments: preambular paragraphs accumulate evidence (recalled treaties, documented problems, stated principles, etc.), and operative paragraphs draw on that evidence to justify a course of action. Our approach is based on two main ideas. First, we use Toulmin's argumentation model (Toulmin, 2003) to break down each resolution into claim, evidence, and warrant, with the warrant explaining how the evidence supports the claim. Second, we define the warrant using evidential reasoning strategies from argumentation and psychology literature. We use four strategies: causal inference (Pearl, 2009), which shows that a premise leads to or supports a directive; corroboration (Godden, 2019; Brem and Rips, 2000), which finds agreement among independent pieces of evidence for the same claim; contrastive reasoning (Lipton, 2013), which points out important differences or opposing views; and triangulation (Breitmayer et al., 1993), which supports a claim using several different methods. We use these strategies as a post-hoc reasoning scaffold for the shared-task relation labels, rather than as a strict one-to-one definition of the official schema.

This scaffold serves a practical purpose: it gives the model a rationale for justifying relation labels rather than choosing them arbitrarily. When the model names a strategy, it commits to a specific logical relationship that can be associated with the paragraph text. At the same time, we treat the strategy-label connection as a prompting device, not as proof that the four strategies perfectly align with the task labels.

The pipeline processes one resolution at a time in two sequential steps. Before either step runs, we build a semantic index over all paragraph texts using `sentence-transformers` (`all-MiniLM-L6-v2`) (Reimers and Gurevych, 2019), which we use in Step 2 to narrow the candidate set. Figure 1 summarizes this separation between candidate discovery and relation interpretation. The first stage is prognostic: semantic similarity estimates which paragraphs are likely to belong together, without yet deciding whether the connection is *supporting*, *complemental*, *contradictive*, or *modifying*. This stage produces a small set of candidate paragraph pairs for the target paragraph. The second stage is diagnostic: the LLM receives those retrieved pairs, the task schema, and the Evident Framework, and then explains how each pair is argumentatively connected. The framework is therefore used to expose the type of relationship and generate the evidence-grounded `think` trace, not to retrieve the pair itself.

The shared task provides relation labels, but the labels alone do not tell the model what reasoning test to apply. The Evident Framework attaches a question to each label. For *supporting*, the model asks whether one paragraph gives a reason, premise, or justification for another. For *comple-*

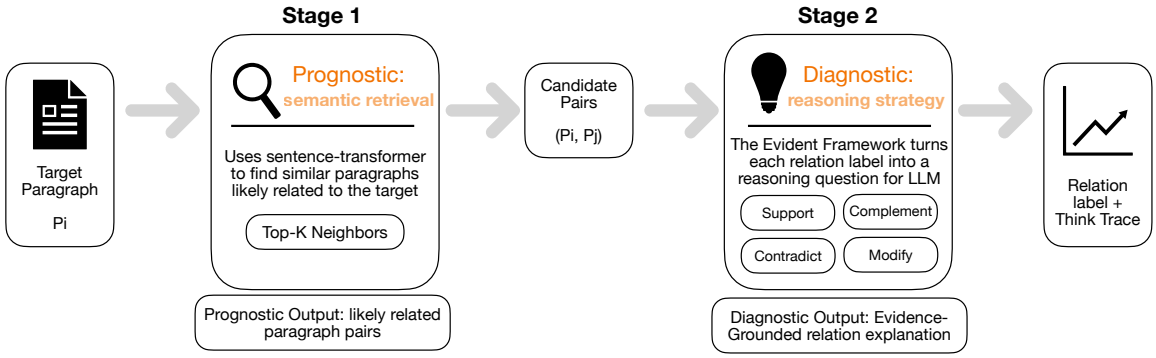

Figure 1: **The two-stage Evident framework** for uncovering argumentative links between resolution paragraphs. Given a target paragraph P_i, the *Prognostic* stage uses a sentence-transformer to retrieve its top-K semantically similar neighbors, producing candidate pairs (P_i, P_j). The *Diagnostic* stage then reformulates each relation label (SUPPORT, COMPLEMENT, CONTRADICT, MODIFY) as a reasoning question posed to an LLM, yielding a predicted label together with a think-trace that grounds the decision in textual evidence.

mental, it asks whether two paragraphs reinforce the same theme through different evidence. For *contradictive*, it asks whether one paragraph introduces opposition, tension, or an incompatible position. For *modifying*, it asks whether one paragraph narrows, conditions, or qualifies the scope of another. This makes the transition from free-text reasoning to task labels smoother: the model is not only selecting a label, but also explaining the corresponding inference using quoted evidence.

Step 1: Classification. For each paragraph, we provide a sliding context window of roughly 21 surrounding paragraphs (indices $[\max(0, i{-}8),\ \min(N, i{+}13)]$) and ask the model to assign a type and tags. To keep the model honest, the prompt requires a four-part reasoning trace (Wei et al., 2022): it must quote the exact opening keyword that signals the paragraph type, name the specific phrase in the text that justifies each tag, identify the dominant reasoning strategy, and explain what role the paragraph plays in the resolution's argument. Paragraphs that contain no substantive claim—document headers, date stamps, institutional name lines—are flagged here and skipped in Step 2.

Step 2: Relation prediction. For each argumentative paragraph, we retrieve its top-15 semantic neighbours from the index (minimum cosine similarity 0.03) and filter out any that Step 1 flagged as non-argumentative. This retrieval step identifies candidate paragraph pairs before any reasoning strategy is assigned. The model is then given the source paragraph alongside its Step 1 metadata and the filtered candidates, and asked to predict which pairs are argumentatively connected and what the relation type is. The Evident Framework is therefore applied after candidate selection: it helps the model explain how a retrieved pair is connected, but it is not the mechanism used to retrieve the pair. Each predicted relation requires a five-part trace: a quoted phrase from the source, a quoted phrase from the target, the relation label with a brief justification, the reasoning strategy and its logical mechanism, and one relation type that was considered and rejected. This last requirement—naming what was ruled out—turned out to be important in practice; without it, the model defaulted to *complemental* for most pairs because two thematically identical paragraphs look superficially similar.

Output. After both steps complete, we combine the Step 1 and Step 2 reasoning traces into a single `think` field per paragraph, with relation traces tagged by their target paragraph number (e.g., `[→ para 5] ...`). A short document-level summary is written to `METADATA.structure.think`. Relation labels are stored in `matched_pars` as `{"target_para_number": [relation_types]}`, matching the official submission schema.

4 Experimental Setup

This section describes the generation models used for inference and the evaluation protocol used to assess the quality of the reasoning traces.

4.1 Generation Models

To assess how well the Evident Framework transfers across model families, we run experiments with two open-source models of comparable scale, both evaluated on the same NVIDIA RTX 4080 (16 GB VRAM) using identical prompts and the same two-step pipeline. The primary model is **Qwen3-8B-GGUF** (Team, 2025), loaded in

Q8_0 quantisation via `llama-cpp-python`; it fits entirely in VRAM with no CPU offloading, and we prepend `/no_think` to suppress its internal chain-of-thought, stripping any residual `<think>...</think>` blocks before parsing. We additionally evaluate **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024) (Meta) under the same hardware and prompt configuration. Across both models, generation uses `temperature=0.1` with a repeat penalty of 1.1 to discourage looping, and token limits of 2,048 for Step 1 and 4,096 for Step 2.

4.2 Internal Diagnostic Evaluation

Because the organizers did not release gold labels for the held-out test set, we could not compute official F1 ourselves or reproduce the official shared-task evaluation. We therefore use **Claude Sonnet 4.6** only for an internal diagnostic evaluation of reasoning-trace quality (Zheng et al., 2023; Liu et al., 2023) via the Anthropic API. The judge sees only the `think` field—no paragraph labels, no task schema, no information about which model generated the text. This setup is not intended to replace the official metric, which uses the task's own evaluation protocol and judge model. Following the rubric-based evaluation approach of Liu et al. (2023), each trace is scored from 0 to 100 across four criteria worth 25 points each, each grounded in an established evaluation principle:

Specificity measures whether the trace quotes actual phrases from the source paragraph rather than paraphrasing vaguely. This mirrors faithfulness evaluation in NLG (Maynez et al., 2020), where ungrounded claims are treated as hallucinations regardless of surface plausibility.

Correctness measures whether the type, tag, and relation decisions are plausible and whether the model considered and rejected alternatives. This aligns with standard argumentation mining evaluation practice (Stab and Gurevych, 2014), extended here to assess the reasoning behind the label rather than the label alone.

Depth & Significance measures whether the trace explains the paragraphś structural role in the resolutionś argument, not just its surface content. This draws on argumentation quality research (Wachsmuth et al., 2017), which identifies argument depth—explaining *why* a claim matters, not just *what* it says—as a primary quality dimension distinct from correctness.

Strategy Precision measures whether the named reasoning strategy is correctly applied and its logical mechanism is shown rather than merely stated. This is grounded in argumentation scheme theory (Walton and Reed, 2005; Toulmin, 2003), which holds that the validity of an inference depends on correctly identifying its underlying scheme and satisfying its critical questions.

Relation grouping and penalization. The four relation types are not equally distinct from one another. *Complemental* (Corroboration) and *modifying* (Triangulation) share a common argumentative function: both add nuance or supporting context without directly justifying or opposing a claim. By contrast, *supporting* (Causal) encodes direct justification and *contradictive* (Contrastive) encodes direct opposition — each maximally distinct from the other and from the complemental/modifying group. For this internal analysis, we therefore apply a hierarchical penalization scheme: cross-group confusions are penalized more heavily than within-group confusions. Predicting *complemental* when the gold label is *modifying* is treated as a softer error than predicting *supporting* when the gold label is *contradictive*. This choice makes the diagnostic rubric framework-dependent and should not be interpreted as equivalent to the official relation-label F1, where such confusions remain full label errors.

To avoid scores being dominated by one relation type—*complemental* tends to be over-predicted—we sample think fields using stratified sampling across seven strata: preambular paragraphs, operative paragraphs, and one stratum per relation type. Each stratum gets an equal share of a 20-sample budget per document.

5 Results

Table 1 shows our internal diagnostic LLM-judge scores on both splits. Because the organisers have not released gold annotations for the test set, we cannot compute official F1 ourselves; these scores therefore assess reasoning-trace quality rather than official task performance. The four criteria give a clearer picture of where the generated reasoning traces are strong and where they remain fragile. They should be read as framework-dependent diagnostics: a model can produce coherent Evident-Framework-style explanations while still assigning some relation labels that differ from the official gold annotations.

This distinction also helps explain the gap between pair identification and exact relation labeling. Candidate pairs are first proposed by semantic

Criterion	Qwen3-8B-GGUF	Llama-3.1-8B -Instruct
Specificity	83	82
Strategy Precision	80	79
Correctness	78	76
Depth & Significance	75	71
Overall (Train)	84	81
Overall (Test)	79	77

Table 1: Internal diagnostic LLM-as-a-Judge scores (0–100) per criterion and per model, evaluated using Claude Sonnet 4.6 as the judge. Per-criterion scores are on the test set; overall scores shown for both splits.

similarity, so the system can retrieve paragraphs that plausibly belong together even when the later diagnostic step chooses the wrong relation type. The problem is most visible for *complemental* and *modifying*, whose boundary is subtle: both involve adding information to an existing theme, but *modifying* additionally changes or qualifies that theme.

Specificity. This was consistently the strongest criterion. Requiring the model to quote exact phrases in every sentence of the reasoning trace turns out to be an effective forcing function: the model rarely falls back on generic statements when it knows the judge is looking for verbatim evidence.

Strategy Precision. Adding litmus-test instructions for each relation type made a noticeable difference. Early runs without these instructions over-predicted *complemental*, because any two thematically identical paragraphs superficially look like mutual reinforcement. Once we required the model to name one relation it had ruled out and explain why, predictions became more discriminating. However, this also shows the limitation of the framework: committing to a reasoning strategy can over-regularize fine-grained label decisions, like, when distinguishing *complemental* from *modifying*.

Depth & Significance. This showed the most variance. Short resolutions with only a few preambular paragraphs gave the model little to work with in explaining structural roles, resulting in lower scores. Longer resolutions with layered preambular chains scored better, since the sliding context window gave the model enough surrounding evidence to make meaningful structural claims.

Correctness. Correctness scores were moderate overall, which is expected: without gold labels, the judge is estimating plausibility rather than measuring accuracy. Tag assignment was the weakest sub-component, partly because the 126-code vocabulary is large and many codes are closely related.

6 Limitations

The main constraint on our evaluation is the absence of gold annotations. The judge score is a useful diagnostic proxy for trace quality, but it is not the same as measuring whether the predicted labels are actually correct under the official schema. A secondary limitation is that Qwen3-8B processes each paragraph independently, which means early classification decisions cannot be revised once later paragraphs provide additional context. Finally, Q8_0 quantisation may introduce small degradations on long structured outputs, though we did not observe obvious failure modes in practice.

7 Conclusion

We presented the POINTERS system for the UZH ArgMining 2026 Shared Task, treating resolution annotation as a reasoning problem rather than a flat classification task. Structured evidential reasoning probes embedded in every prompt force the model to commit to textual evidence, name the logical mechanism, and rule out alternatives—making decisions falsifiable rather than merely fluent. Running on a local Qwen3-8B model, the system produces grounded reasoning traces under our internal diagnostic evaluation. At the same time, our results show that an interpretable reasoning scaffold should not be treated as a perfect substitute for the official relation schema: semantic retrieval can identify plausible paragraph pairs, while the subsequent strategy-based diagnostic step may still confuse fine-grained labels such as *complemental* and *modifying*. The fine-tuning on gold-labelled data once released would likely improve label accuracy, decoupling pair retrieval from relation labeling may reduce framework-induced label errors, and incorporating the French source text may recover argumentative cues lost in translation (Palau and Moens, 2009; Cabrio and Villata, 2018).

8 Acknowledgment

This work is supported in part by the PROTECT project, awarded by the U.S. Department of Energy's (DOE) Office of Cybersecurity, Energy Security, and Emergency Response (CESER) to Pacific Northwest National Laboratory (PNNL) through solicitation RC-40125b-2023; and by the Collaborative Research, Innovation and Strategic Partnerships (CRISP) grant at NJIT.

References

Bonnie J. Breitmayer, Lioness Ayres, and Kathleen A. Knafl. 1993. Triangulation in qualitative research: Evaluation of completeness and confirmation purposes. *Image: The Journal of Nursing Scholarship*, 25(3):237–243.

Sarah K. Brem and Lance J. Rips. 2000. Explanation and evidence in informal argument. *Cognitive Science*, 24(4):573–604.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Giuseppina Scotto Di Carlo. 2013. *Vagueness as a political strategy: Weasel words in security council resolutions relating to the second gulf war*. Cambridge Scholars Publishing.

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.

David Godden. 2019. Corroboration: Sensitivity, safety, and explanation. *Acta Analytica*, 34(1):15–38.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational linguistics*, 43(1):125–179.

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.

Peter Lipton. 2013. Inference to the best explanation. In *The Routledge Companion to Philosophy of Science*, pages 225–234. Routledge.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Judea Pearl. 2009. *Causality*. Cambridge University Press.

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 and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Manfred Stede, Jodi Schneider, and Graeme Hirst. 2019. *Argumentation mining*. Springer.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Stephen E. Toulmin. 2003. *The Uses of Argument*, updated edition. Cambridge University Press.

Sahaj Vaidya and Aritra Dasgupta. 2020. Knowing what to look for: A fact-evidence reasoning framework for decoding communicative visualization. In *2020 IEEE Visualization Conference (VIS)*, pages 231–235. IEEE.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Douglas Walton and Chris A Reed. 2005. Argumentation schemes and enthymemes. *Synthese*, 145(3):339–370.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.