Overview of the UZH Shared Task 2026 on Reconstructing the Reasoning in United Nations Resolutions

Anastassia Shaitarova[†UZH IFE] **Yingqiang Gao**[UZH LiRI]
Fatma-Zohra Rezkellah[UZH CL] **Reto Gubelmann**[UZH CL, DSI] **Patrick Montjouridès**[†UZH IFE]
[UZH IFE] Institute of Education, University of Zurich, Switzerland
[UZH LiRI] Linguistic Research Infrastructure, University of Zurich, Switzerland
[UZH CL] Department of Computational Linguistics, University of Zurich, Switzerland
[UZH DSI] Digital Society Initiative, University of Zurich, Switzerland
{firstname.lastname}@uzh.ch
https://shared-task-argmining.linguistik.uzh.ch/

Abstract

This paper presents the UZH Shared Task at the 13th Workshop on Argument Mining and Reasoning, co-located with ACL 2026, which focuses on reconstructing argumentative structure in highly formal legal-political texts, namely United Nations resolutions and recommendations. The shared task addresses the challenge of recovering paragraph-level reasoning patterns from the fairly formulaic structure of international decision-making records. It comprises two subtasks: (1) paragraph classification, where systems identify paragraph type (preambular or operative) and assign one or more thematic tags, and (2) argumentative relation prediction, where systems infer links between paragraphs and label them with relation types such as supporting, contradictive, complemental, and modifying. The data is provided in French, together with English translations. It includes a test set of resolutions and recommendations from the UNESCO International Conference on Education, annotated at paragraph level, as well as unlabeled training resolutions from the United Nations. The task restricts submissions to open-weight language models with at most 8 billion parameters and requires demonstration of models' reasoning capabilities. By launching this task, we aim to advance research on argument mining in formal institutional discourse, particularly in multilingual, policy-oriented documents.

Dataset Code

1 Introduction

Resolutions adopted by United Nations (UN) bodies represent a distinct and underexplored genre for argument mining. These "formal expressions of the opinion or will of United Nations organs" (United Nations, 1983, p.167) can be seen as structured argumentative texts: they encode negotiated positions, implicit premises, and carefully structured conclusions representing the highest level of intergovernmental consensus on international issues (Bernstein, 2011). Achieving political consensus on global education values can take decades of negotiation before such agreements are formalized in intergovernmental declarations, resolutions, or recommendations that form a cornerstone of intergovernmental cooperation towards shared global objectives (e.g., Sachs-Israel, 2016).

Although most of these documents are not legally binding, every word is carefully weighed, scrutinized, and validated. Structurally, resolutions often contain preambular and operative sections. The preambular part lays out the considerations based on which an action is undertaken, an opinion is expressed, or a directive issued, while the operative part articulates the actual position and recommendations adopted by UN bodies' Member States (United Nations, 2025).

Beyond this formal structure, UN resolutions exhibit idiosyncratic argumentative characteristics: deliberate formulations that are simultaneously open-ended and precisely calibrated, and a self-referential intertextual ecosystem through the systematic citation of prior resolutions (Scotto di Carlo, 2013, 2017). Given the influence of United Nations texts on education systems worldwide, shaping for instance visions of educational justice and fairness (Montjouridès, 2022), examining the underlying reasoning structures both within individual resolutions and across documents over time can illuminate questions of interest to scholars in international relations, education, digital humanities, and computational social science, among others.

Motivated by the lack of research in this area, we

[†]Corresponding authors.

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 107–118

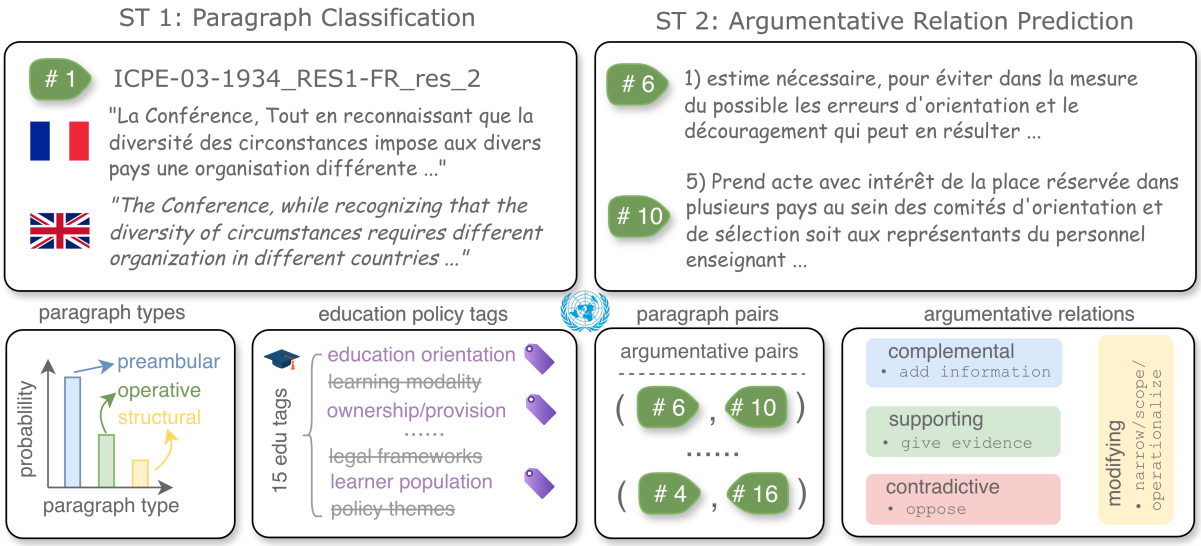

Figure 1: Overview of the UZH Shared Task 2026 on Reconstructing the Reasoning in United Nations Resolutions.

launched the UZH Shared Task, *Reconstructing the Reasoning in United Nations Resolutions*, as part of the 13th Workshop on Argument Mining and Reasoning. We tasked participants with paragraph-level structural classification into preambular and operative sections, assignment of educational policy tags, as well as identification of argumentative relations, all within resolutions and recommendations adopted at UNESCO's International Bureau of Education (IBE) International Conferences on Education (1934–2008).

In line with the workshop theme "*understanding and evaluating arguments in both human and machine reasoning*," we required participants to enable Chain-of-Thought (CoT; Wei et al., 2022) or thinking mode in their models and submit the produced reasoning chains. Furthermore, all systems had to rely exclusively on open-weight models with at most 8 billion parameters, probing whether current small open-source LLMs can move beyond surface pattern matching to recover the implicit reasoning structures of institutional argumentation.

Eight teams participated with one team submitting two runs. The evaluation combines automatic F1 metrics with an LLM-as-a-judge protocol, both applied against a silver-standard ground truth validated against human domain expert annotations. Results across submissions show that (i) the classification of paragraphs into preambular and operative types is largely solved by rule-based approaches and fine-tuned models; (ii) thematic tag annotation remains a difficult open problem, particularly for rare tags; (iii) all teams systematically over-predict argumentative pairs, and relation label classification is bounded by genuine annotation ambiguity ($\kappa = 0.540$), with current systems reaching 65% of the human ceiling; (iv) F1-based and LLM-as-a-judge rankings diverge substantially, indicating that output accuracy and reasoning quality are complementary but not equivalent objectives; (v) self-evaluation against a proprietary LLM cannot be assumed to predict performance against a different judge model.

2 Task Formulation

The UZH Shared Task consists of two subtasks covering paragraph classification and argumentative relation prediction (Figure 1), both requiring familiarity with the domain and with the structural conventions of institutional resolutions.

2.1 Subtask 1 (ST 1)

(a) Paragraph type classification. The first component of ST1 is to assign a preambular or operative type label to the paragraphs of the resolution. Most resolutions follow a strict two-block structure, starting with a preambular block explaining the rationale, context, or basis for the measures called for. Preambular paragraphs introduce, contextualize, or justify the resolution and typically begin with a present participle (e.g. *Acknowledging*, *Recalling*, *Noting that*, *Having reviewed*).

The preambular section is usually followed by operative paragraphs declaring what the conference decides, recommends, or calls for. Operative paragraphs enact the resolution's decisions and often begin with present-tense action verbs (e.g. *Adopts*, *Endorses*, *Notes with approval*, *Urges*).

These blocks never interleave, and once the operative section begins, it does not revert to preambular paragraphs. This yields three paragraph classes in total: preambular, operative, and a rare structural class covering non-argumentative elements such as section headers, enumeration openers, or annexes.

(b) Education policy tag annotation. The second component of ST 1 is to label each paragraph with one or more thematic tags from a fixed ontology covering 15 education policy dimensions: ISCED (International Standard Classification of Education; UNESCO (2012)) education levels, education orientation, learning modality, ownership/provision, teachers, infrastructure and resources, curriculum, pedagogy and assessment, subject domains, cross-cutting themes and skills, policy themes, education system monitoring and evaluation, legal frameworks, stakeholder focus, and learner population.

2.2 Subtask 2 (ST 2)

Subtask 2 reflects the core argument mining challenge and carries greater weight in the final ranking, as argumentative relation prediction is both more difficult and more central to the shared task's goals.

(a) Argumentative relationship detection. Given a document, participants must identify which pairs of paragraphs stand in an argumentative relation. By convention, only later paragraphs (higher paragraph numbers) hold relations pointing back to earlier ones; pairs are treated as unordered for evaluation.

(b) Argumentative relationship labeling. Each identified pair must be assigned one of four relation types: **complemental** (two paragraphs make independent parallel contributions to the same broader argumentative goal, without one directly reacting to the specific claim of the other), **modifying** (one paragraph narrows, scopes, or operationalizes another), **supporting** (a principle and a piece of evidence reinforce each other, or evidence provides justification for a proposal), or **contradictive** (one paragraph introduces tension, a qualification, or a contradiction with respect to another).

2.3 Evaluation and Ranking

The official ranking proceeds in two stages.

F1-based combined ranking. We use scikit-learn's `classification_report` to produce F1 scores for each component. Given the class imbalance in both ST1 components, we evaluate them using macro-averaged F1. Evaluation of the ST2 components uses F1 score, weighted by label frequency in the test set over the four relation types on correctly identified pairs only. Ranks are assigned per component and averaged within each subtask. The final rank combines subtask ranks with ST2 carrying greater weight, reflecting that argumentative relation prediction is the more central challenge of the task (Equation 1). Equal scores share the same rank.

$$\text{rank}_{\text{F1}} = 0.4 \times \text{rank}_{\text{ST1}} + 0.6 \times \text{rank}_{\text{ST2}} \quad (1)$$

LLM-as-a-Judge evaluation. In addition to the automatic F1 evaluation, all submissions are assessed by an LLM-as-a-judge protocol using an open-weight LLM with a fixed prompt on a 1–100 scale, yielding an independent quality ranking (see Section 5.4).

Final ranking. The final rank combines the F1-based ranking and the LLM-as-a-judge rank, with F1 carrying greater weight (0.6) as the primary accountability signal against shared ground truth, while LLM-as-a-judge contributes a complementary quality dimension (Equation 2)

$$\text{rank}_{\text{final}} = 0.6 \times \text{rank}_{\text{F1}} + 0.4 \times \text{rank}_{\text{LLM-judge}}. \quad (2)$$

3 Data

3.1 Training Data

Participants were provided with 2,695 UN resolutions from the UN-RES corpus (Gao et al., 2025), originally in French with machine-generated English translations produced using the Helsinki-NLP `opus-mt-fr-en` model (Tiedemann and Thottingal, 2020). Unlike a conventional labeled training set, this corpus carries no paragraph-level annotations and was provided as auxiliary unlabeled data for unsupervised use, such as domain familiarization, in-context learning (ICL; Brown et al. (2020)), retrieval-augmented generation (RAG; Lewis et al. (2020)), or other LLM-based techniques. The corpus is distributed under a restricted UN license; participants agreed not to redistribute it publicly.

3.2 Test Data

The test set consists of 45 parsed documents (92 individual resolutions and recommendations) from the UNESCO IBE International Conferences on Education (1934–2008). Documents are provided in French with machine-generated English translations using `gpt-4.1-mini`.

3.3 Ground Truth Construction

Ground truth annotations follow a *silver-standard* approach: LLM-generated labels developed iteratively against a human-annotated validation sample drawn from the test corpus. Table 1 summarises the dataset statistics.

	Train	Test
Documents	2,695	45
Resolutions	—	92
Human gold (validation subsets)		
Type-annotated paragraphs	—	715
Expert (11 docs)	—	363
Second annotator (7 docs)	—	352
Relation-annotated pairs	—	300
Tag-reviewed paragraphs	—	178

Table 1: Dataset statistics. Training data is unannotated. Human gold figures refer to the validation subsets used for secondary evaluation.

Paragraph types. The silver-standard type labels were generated by `gpt-5.4-mini` at temperature 0. Validation of the model output against expert human annotations on 363 paragraphs (11 documents) yielded 97.0% accuracy (preambular F1: 0.96, operative F1: 0.98, structural F1: 0.89); a second independent validation on 352 paragraphs (7 different documents) yielded 97.4% accuracy, giving a combined human-validated set of 715 paragraphs across 18 documents.

Thematic tags. Tags were generated by both `gpt-5.4` and `claude-opus-4.6` using a shared prompt. The silver-standard gold uses the *intersection* of both models' predictions as a conservative estimate. A domain expert approved the intersection labels on 178 paragraphs (the same paragraphs sampled for relation annotation).

Argumentative relations. Relations were annotated by `gpt-5.4` using the best-performing prompt from a six-version iterative development process at temperature 0. The gold standard consists of 300 paragraph pairs annotated in two rounds by the same expert annotator; these pairs were also used throughout prompt development, so the evaluation is not fully independent of the development process. The pairs were not sampled randomly: they were stratified across cases where an initial two-model run agreed and disagreed, to ensure coverage of the full annotation space.

Round 2 involved blind re-annotation of 233 pairs (without access to round-1 labels), resulting in 64 label changes. The intra-annotator agreement on these 233 re-annotated pairs is $\kappa = 0.540$, which we treat as the honest human ceiling; agreement on all 300 pairs inflates to $\kappa = 0.664$ due to unchanged pairs. These values place the task in the range reported for argument relation annotation in other domains (Lauscher et al., 2018; Lawrence and Reed, 2019).

The best model achieves $\kappa = 0.354$ against the round-2 gold which amounts to 65% of the honest human ceiling. Extended-thinking mode with `claude-opus-4.6` ($\kappa = 0.344$) does not improve over standard GPT inference ($\kappa = 0.354$), suggesting that additional reasoning budget does not resolve the inherent label ambiguity.

The relation gold standard is highly imbalanced (Table 6): modifying (55%), complemental (26%), none (13%), supporting (5%), and contradictive (0%). The main human disagreement is between complemental and modifying, which accounts for 28% of label changes in the re-annotation round (see Appendix A).

4 Overview of the Submissions

All eight teams follow a pipeline-based design decomposing the task into sequential modules. Systems differ primarily in how they incorporate reasoning: several use chain-of-thought prompting or multi-agent debate, while others rely on retrieval-augmented generation. Six of eight teams used Qwen2.5-7B or Qwen3-8B as their primary generation model, consistent with the task's constraints. Figure 2 provides an overview of the methodological techniques adopted across submissions.

LLM-Instruct (Huu Vu Tran et al., 2026)**.** The LLM-Instruct team submitted two runs of the same system. The system frames the task as a retrieval-

and-decoding problem. For thematic tagging, candidate labels are first retrieved by dense similarity, then narrowed by semantic dimension and validated against a strict closed set, preventing hallucinated tags. For uncertain paragraphs, the system triggers a three-agent debate (three different open-weight 8B models) and applies a constrained aggregator to resolve disagreements. Relation candidates are generated by locality and embedding similarity, filtered by a confidence threshold.

Argchestrators (Grecu et al., 2026)**.** Argchestrators proposes a modular hybrid architecture with three distinct components. For type classification, structural and lexical heuristics handle clear cases and only edge cases are forwarded to a zero-shot LLM call. For thematic tagging, a hierarchical multi-agent debate is used: an Expansionist agent (biased towards recall) and a Skeptic agent (biased towards precision) argue over high-level ontological dimensions, supervised by an Orchestrator, before a final dimension-specific pass selects low-level tags. For relation prediction, asymmetric distance-decay retrieval models the backward-referencing structure of UN texts, penalizing forward-looking candidates. Argchestrators achieves the best type F1 (0.936) and relation label F1 (0.440) in the silver evaluation and tied third with HybridArguer.

Prompteam (Khandelwal and Bhardwaj, 2026)**.** Prompteam combines RAG with engineering optimisations for constrained GPU environments. For Subtask 1, dense embeddings retrieve candidate tags and few-shot examples before an LLM produces a structured chain-of-thought output. For Subtask 2, a sliding-window cosine pre-filter reduces the quadratic candidate space to near-linear, substantially cutting LLM calls. The system was parallelized across multiple sessions with atomic checkpointing to achieve full test-set coverage under strict GPU time limits. Despite weaker F1 scores, Prompteam ranks second overall on the strength of its top LLM-judge score (rank 1), indicating that its reasoning traces were judged the most coherent and well-structured.

POINTERS (Sen et al., 2026)**.** POINTERS treats resolution annotation as a structured reasoning task rather than a classification problem. Grounded in the Evident Framework, the system maps the four relation types onto four explicit reasoning strategies (Causal, Corroboration, Contrastive, and Triangulation) and requires the model to name the strategy, quote supporting evidence, and explicitly rule out alternatives for every predicted relation. Running entirely locally on a consumer GPU without cloud API calls, the system produces interpretable five-part reasoning traces. The authors additionally conducted an independent judge evaluation using Claude Sonnet 4.6, reporting scores of 81/100 on training and 77/100 on the test set; these are not directly comparable to the official shared task judge scores, which use a different model and scoring criteria. POINTERS achieves the best tag scores by a notable margin (micro-F1 0.459, macro-F1 0.357) and the best pair F1 (0.330) in the silver evaluation, but ranks last in the official LLM-judge evaluation with a final score of 26.16, driven by particularly low logical and dialectical scores.

TypeCoT (Kumari et al., 2026)**.** TypeCoT introduces a type-informed chain-of-thought approach where structural predictions from Subtask 1 are explicitly reused as constraints in Subtask 2. For type classification a LoRA-fine-tuned Qwen2.5-7B-Instruct model is applied; tag assignment is performed dimension-by-dimension, with a separate inference call for each of the 15 ontological dimensions. For relation prediction, the predicted paragraph type acts as a constrained prior that shapes the candidate generation process, guiding the model to reason about structural compatibility before predicting link labels. A multi-pass recovery pipeline handles context overflow on long documents. TypeCoT achieves strong type F1 (0.913) but relatively low tag and pair scores, and ranks eighth overall.

HybridArguer (Bhargava, 2026)**.** HybridArguer proposes a modular pipeline decomposing the task into document-level and paragraph-level inference steps. For type classification, a zero-shot Qwen3-8B call in thinking mode processes the full document with two-pass majority-vote self-consistency. For thematic tagging, multilingual-e5-large embeddings over French and English representations retrieve top tag candidates via kNN cosine similarity, passed to a paragraph-level LLM for multi-label selection. Relation candidates are selected analogously by embedding similarity, supplemented

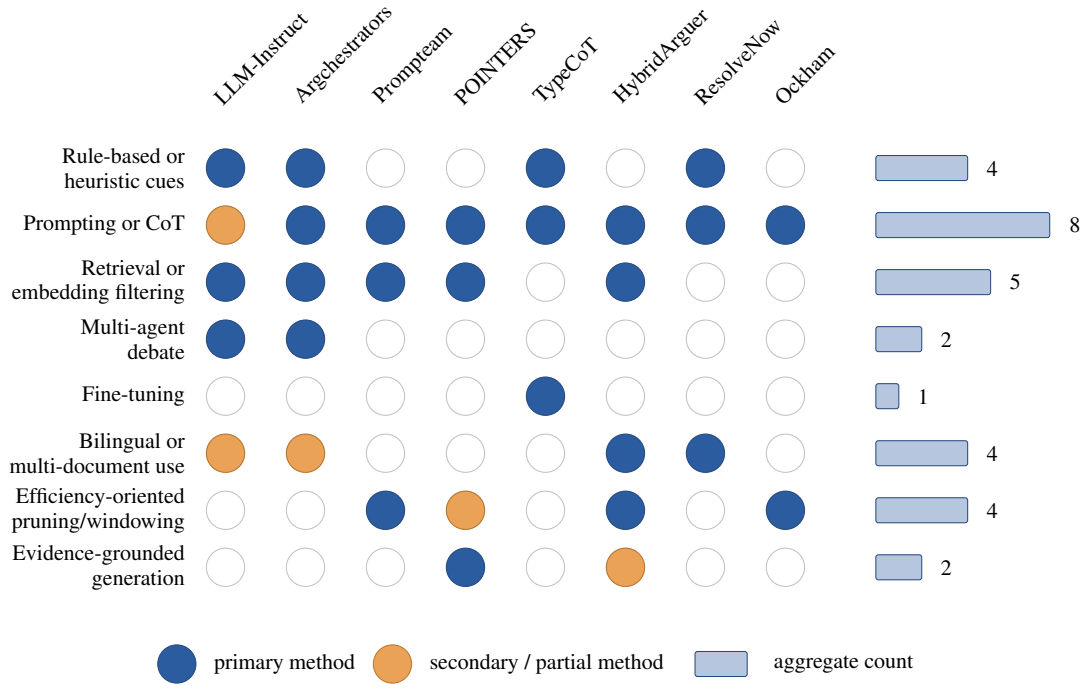

Figure 2: Overview of methodological techniques used by participating teams.

by the three immediately following paragraphs; the same call predicts relation existence and type with confidence scoring. Iterative prompting with corrective feedback stabilizes structured outputs throughout. The system achieves type macro F1 of 0.891 and relation label F1 of 0.389 in the silver evaluation, with notably high pair recall (0.713), and tied third overall.

ResolveNow (Gupta et al., 2026)**.** ResolveNow handles type classification entirely by deterministic pattern matching over 100 French and 60 English lexical patterns, spending no LLM budget on what the authors treat as a solved structural problem given the formulaic drafting conventions of UN texts. Thematic tagging is delegated to a reasoning-augmented 8B model that receives the full 141-tag taxonomy in a single prompt, batching three paragraphs per call. ResolveNow achieves strong type F1 (0.910) but submitted no predictions for Subtask 2, resulting in zero pair and relation scores.

Ockham. Ockham is built around a quantized Llama-3.1-8B-Instruct model and focuses on computational efficiency under the 8B constraint. Its central mechanism is Semantic Entropy Pruning, which uses token-level entropy to identify and discard low-information context before inference; entropy thresholds were calibrated through empirical analysis of the training corpus. For relation prediction, attention is restricted to a sliding window of three immediately preceding paragraphs rather than the full document. This aggressive context management reduces processing cost substantially but at the expense of recall: Ockham achieves the lowest type F1 (0.445), suggesting that the entropy-based pruning may also discard argumentatively relevant content.

5 Results and Discussion

5.1 Silver-Standard Evaluation

Team	Type mF1	Tag μF1	Tag mF1	Pair P	Pair R	Pair F1	Rel F1
Argchestrators	**0.936**	0.327	0.285	0.119	0.740	0.206	**0.440**
HybridArguer	0.891	0.380	0.224	0.173	0.713	0.279	0.389
LLM-Instruct	0.815	0.396	0.294	0.205	0.748	0.322	0.366
LLM-Instruct-2	0.891	0.329	0.254	0.162	0.270	0.202	0.350
Ockham	0.445	0.205	0.045	0.194	0.569	0.289	0.328
POINTERS	0.762	**0.459**	**0.357**	**0.208**	**0.796**	**0.330**	0.286
Prompteam	0.587	0.226	0.169	0.136	0.611	0.222	0.413
ResolveNow	0.910	0.344	0.236	—	—	—	—
TypeCoT	0.913	0.280	0.278	0.072	0.401	0.123	0.329

Table 2: Results on the silver-standard test set. μF1 denotes micro-averaged F1; mF1 denotes macro-averaged F1. Pair P/R/F1 measure pair identification (unordered); Rel F1 is weighted F1 over relation labels on correctly identified pairs only.

Table 2 reports raw scores on the full silver-standard test set. Argchestrators achieves the best type macro F1 (0.936) and relation label F1 (0.440); POINTERS leads on all tag metrics (micro-F1 0.459, macro-F1 0.357) and pair F1 (0.330). In five submissions, type macro F1 exceeds 0.89. All

teams systematically over-predict argumentative pairs, with pair F1 remaining moderate despite high recall across all architectural strategies (Section 6).

5.2 Human Gold Evaluation

Table 7 reports results against the human-annotated validation samples. Type macro F1 scores shift modestly and rankings are broadly preserved. Within the 300-pair annotated sample, pair identification precision is high across all teams (0.835–0.947), and pair F1 (0.338–0.838) exceeds the corresponding silver values. This reflects the stratified sampling design: the gold pairs were selected around model predictions, so systems find most of them by construction, and with only 39 explicit negatives in the sample there is limited scope for false positives. Rankings are broadly preserved: POINTERS leads pair identification (0.838), followed by LLM-Instruct (0.770) and Argchestrators (0.763). Relation label F1 is modestly lower than in the silver evaluation, with rankings broadly stable: Argchestrators retains the lead (0.380) and LLM-Instruct's second run posts the second-highest relation F1 (0.372). The broad stability of rankings across both evaluations suggests that the silver-standard labels are a reasonable proxy for human judgment, particularly for type classification and relation labeling (see Table 7 in Appendix A).

5.3 Combined F1 Ranking

Team	ST1	ST2	Score	F1 Rank
LLM-Instruct	3.5	1	2.0	1
Argchestrators	1	3	2.2	2
POINTERS	3.5	5	4.4	3
HybridArguer	7	3	4.6	4
Prompteam	8	3	5.0	5
TypeCoT	2	8	5.6	6
LLM-Instruct-2	5.5	7	6.4	7
Ockham	9	6	7.2	8
ResolveNow	5.5	9	7.6	9

Table 3: F1-based combined ranking of the two subtasks (ST1 and ST2). Score $= 0.4 \times \text{rank}_{\text{ST1}} + 0.6 \times \text{rank}_{\text{ST2}}$; equal scores share the same rank. ST1: average of type macro F1 and tag macro F1 component ranks. ST2: average of pair F1 and relation label F1 component ranks. Ties are handled using average ranks, which may result in fractional values (e.g., 3.5); these are resolved in the final weighted ranking.

Table 3 shows the official F1-based ranking across all submissions. LLM-Instruct ranks first overall, leading in ST2 (rank1) while tied for third in ST1. Argchestrators ranks second, with the best ST1 performance (rank1) but a weaker ST2 (rank3). POINTERS rises to third, leveraging the best tag macro F1 (0.357) despite weaker relation prediction. The middle positions reflect different subtask trade-offs: TypeCoT's strong ST1 rank (2) is offset by a weak ST2 (rank8). LLM-Instruct-2, the second run by the same team, enters at rank7, held back by moderate performance across both subtasks. Ockham (rank8) and ResolveNow (rank9) rank last; ResolveNow's strong type F1 is offset by the absence of Subtask2 predictions.

5.4 LLM-as-a-Judge Evaluation

Team	Log.	Rhet.	Dial.	Final	Rank
Prompteam	89.45	90.30	87.67	89.05	1
ResolveNow	78.53	88.60	58.65	74.73	2
HybridArguer	70.58	82.35	60.00	70.65	3
LLM-Instruct-2	73.87	58.55	68.10	66.82	4
LLM-Instruct	63.79	67.01	59.46	63.50	5
Argchestrators	53.60	70.33	35.42	52.56	6
Ockham	40.51	64.80	29.24	43.86	7
TypeCoT	38.77	57.57	21.48	38.83	8
POINTERS	21.02	42.51	14.06	26.16	9

Table 4: LLM-as-a-judge evaluation results (Gemma-4-E4B-IT), ranked by final score. Log. = logical, Rhet. = rhetorical, Dial. = dialectical.

In addition to automatic F1 metrics, an LLM-as-a-judge protocol was employed, where an open-weight LLM is applied with a fixed prompt to evaluate the quality of each submission's reasoning chains, producing an independent quality ranking complementary to the F1-based evaluation. The judge assesses each thinking chain independently against three criteria from Wachsmuth et al. (2017), as specified by Ivanova and Gubelmann: *logical quality* (cogency and inferential validity), *rhetorical quality* (clarity, conciseness, and persuasive effect), and *dialectical quality* (ability to resolve the argumentative question for a well-informed reader). Each criterion is scored on a 1–100 scale, and the three scores are averaged to report a per-chain final score. Team-level scores are obtained by averaging over all chains across all test documents. The full judge prompt is provided in Appendix B. It should be noted that the judge evaluates the quality of the submitted reasoning chains but cannot verify that these chains causally produced the predictions.

Concretely, we use Gemma-4-E4B-IT (Google, 2026), running on a single A100 80GB GPU, selected for its strong reasoning capabilities at a parameter count suited to high-throughput chain-by-chain evaluation. Reasoning chains across all submissions average approximately 2.28 million input tokens per team, making throughput a critical constraint. Experiments with larger models, namely Qwen3-235B-A22B and Qwen3.5-122B-A10B (Yang et al., 2025), consistently failed to complete within available resources: the low throughput of large Mixture-of-Experts (MoE; Fedus et al. (2022)) models, combined with the scale of chain-by-chain evaluation, proved prohibitively slow, exhausting both memory and time allocations on the cluster.

Results are reported in Table 4. Prompteam ranks first across all three criteria by a substantial margin. ResolveNow and HybridArguer follow, with strong logical and rhetorical scores but weaker dialectical performance. POINTERS ranks last, with notably low logical and dialectical scores.

5.5 Final Rankings

Team	F1 Rank	LLM-Judge	Weighted Sum	Final
LLM-Instruct	1	5	2.6	**1**
Prompteam	5	1	3.4	2
Argchestrators	2	6	3.6	3.5
HybridArguer	4	3	3.6	3.5
POINTERS	3	9	5.4	5
LLM-Instruct-2	7	4	5.8	6
ResolveNow	9	2	6.2	7
TypeCoT	6	8	6.8	8
Ockham	8	7	7.6	9

Table 5: Final rankings. **Final** $= 0.6 \times \text{rank}_{\text{F1}} + 0.4 \times \text{rank}_{\text{LLM-judge}}$; equal scores share the same rank. LLM-as-a-judge ranking described in Section 5.4.

Table 5 shows the final rankings. LLM-Instruct remains the overall winner, retaining its F1-based rank 1 despite a mid-range LLM-judge score (rank 5). Prompteam rises to second place: the best LLM-judge score (rank 1) compensates for its weaker F1 performance (rank 5). Argchestrators and HybridArguer tie at rank 3.5 with identical weighted scores (3.6), albeit via opposite trade-offs Argchestrators leads on F1 (rank 2) while HybridArguer leads on LLM-judge (rank 3). POINTERS drops from third (F1) to fifth overall due to the weakest LLM-judge score (rank 9). LLM-Instruct-2 places sixth, benefiting from a solid LLM-judge score (rank 4) that partially offsets its lower F1 rank (7). ResolveNow climbs to seventh thanks to the second-best LLM-judge score (rank 2), despite submitting no Subtask 2 predictions. TypeCoT and Ockham close the ranking with consistently weaker performance across both dimensions.

6 Analysis

Paragraph type classification. The formulaic structure of UN resolutions makes type classification close to a lexical lookup, and five of nine runs exceed macro F1 0.89. Tellingly, the highest-scoring system sidelines LLMs for this step: Argchestrators (0.936) relies on deterministic bilingual lexical rules, as does ResolveNow (0.910). TypeCoT's fine-tuned LLM (0.913) matches this level, suggesting that fine-tuning is competitive; LLM-Instruct adopted a rule-based heuristic approach after finding it more stable than generative inference. Ockham (0.445) and Prompteam (0.587) are the two outliers: Ockham's entropy-based context pruning likely discards the paragraph-opening verb forms that are the primary type signal, explaining its poor type F1 (0.445); Prompteam delegates type classification to a few-shot LLM, adding unnecessary complexity to a near-solved structural problem. Type classification is the most robustly validated subtask, with $\approx 97\%$ accuracy against two human annotators on 715 paragraphs.

Thematic tag annotation. Tags remain the hardest subtask, and the variation in scores reveals a clear design trap: pre-filtering the 141-tag space by retrieval before generation prevents hallucination but systematically excludes rare labels. Prompteam (micro-F1 0.226), which narrows down to 20 candidates, explicitly acknowledges that rare tags may never enter the candidate set; LLM-Instruct (0.396), using a wider pool of 40 candidates with per-dimension caps to control over-prediction, achieves nearly double the micro-F1. POINTERS leads on both metrics (micro-F1 0.459, macro-F1 0.357) with an evidence-grounded generative approach that requires the model to quote a specific phrase for every tag decision, bypassing retrieval pre-filtering entirely. The near-equal micro and macro-F1 of TypeCoT (0.280 vs. 0.278) is also notable: its dimension-by-dimension prompting (15 calls per paragraph) is the only approach that structurally forces coverage of all ontological dimen-

sions, at substantial inference cost. The consistent micro/macro gap in all other teams confirms that performance concentrates on frequent tags. Expert review of the 178-paragraph gold subset identified 20 paragraphs with genuine ontology gaps, imposing a ceiling below 1.0 even for a perfect system.

Pair identification. All teams systematically over-predict argumentative pairs, with recall consistently exceeding precision regardless of the architectural strategy used: locality windows (Prompteam, Ockham), asymmetric distance-decay (Argchestrators), type-pair structural constraints (TypeCoT), and per-paragraph edge caps (LLM-Instruct). That no approach closes the recall–precision gap suggests the problem is not one of candidate filtering but of the underlying tendency of LLMs to find argumentative connections more liberally than the task requires.

Relation label classification. Relation label F1 is more stable across evaluation conditions. The main source of confusion—both for human annotators and for systems—is the complemental/modifying distinction: 28% of round-2 blind re-annotations changed this label, and several teams explicitly designed mechanisms to combat it (POINTERS required naming a ruled-out alternative; TypeCoT imposed empirical type-pair priors), yet the problem persisted across the board.

Cross-cutting design lessons. Several decisions distinguish system rankings. Explicit encoding of resolution structure pays off: The two teams relying on deterministic lexical rules for type classification achieve the first and third type F1 (Argchestrators and ResolveNow); TypeCoT's fine-tuned LLM is competitive with the top deterministic approaches, while the two outliers (Ockham and Prompteam) both routed type classification through LLM inference rather than exploiting the formulaic structure directly. Furthermore, using sentence-transformer embeddings to pre-filter the tag space is a losing strategy: the teams that pre-filter most aggressively score lowest on tags, while POINTERS, which bypasses retrieval pre-filtering entirely, leads. Finally, POINTERS illustrates the risk of self-evaluation against a mismatched judge: its ClaudeSonnet4.6 self-score of 77/100 contrasts sharply with the official Gemma score of 26.16, confirming that judge scores are not portable across model families.

The provided training corpus of 2,695 UN resolutions was used by only one team in a documented way: Ockham conducted an empirical analysis of entropy profiles on the training corpus to calibrate its entropy pruning thresholds. Despite the bilingual data provision, teams varied considerably in how they exploited the French originals. Most systems appeared to rely primarily on English translations during inference, using French mainly in heuristic or auxiliary roles. ResolveNow made substantive use of the French text through bilingual lexical rule sets for type classification. HybridArguer was the only system to explicitly combine French and English representations at the embedding level for both tag and relation candidate selection.

Remarkably, while the human expert annotations were produced on the French originals, the silver-standard labels were generated from English translations. The broad stability of rankings across these two evaluations suggests that language alignment was not the dominant source of variance across systems. However, because only one team integrated bilingual semantic representations directly within its retrieval pipeline, the current evaluation cannot determine whether more deeply cross-lingual architectures would perform differently. This remains an open question for future shared tasks with more linguistically integrated submissions.

7 Conclusion

We introduced a multi-level argument mining benchmark over a historically significant UNESCO policy corpus. The results point to three design lessons: exploit document structure with deterministic rules rather than unconstrained LLM generation for near-solved subtasks; prefer evidence-grounded generation over retrieval pre-filtering when rare labels matter; and treat systematic pair over-prediction as a fundamental LLM tendency rather than a candidate-filtering problem. Future editions would benefit from a larger human-annotated gold set, an extended ontology addressing documented coverage gaps, and evaluation designs that decouple pair identification from relation labeling. The dual evaluation protocol demonstrates that F1 performance and reasoning quality are independent dimensions: combining both is essential for a full picture of system capability in reasoning-intensive tasks of argument mining.

Limitations

The ground truth for tags and relations is a silver standard produced by LLMs, not independent human annotation. While type annotations are robustly validated ($\approx$97% accuracy against two human annotators), tag and relation quality depend on the model-intersection and single-model strategies respectively. The tag ontology has documented gaps; 20 paragraphs in the expert-reviewed subset could not be adequately labelled with existing categories. The relation gold covers only 300 of the several thousand paragraph pairs in the corpus, sampled non-randomly (stratified by model agreement), and the same 300 pairs were used for both prompt development and final evaluation, introducing a risk of overfitting the annotation methodology to the sample. Finally, the corpus is restricted to a single institutional genre (UNESCO IBE intergovernmental resolutions) and a particular historical period; findings may not generalise to other policy document types or languages.

Acknowledgements

We gratefully acknowledge the DICED* project at the University of Zurich, supported by Swiss Open Research Data Grants (CHORD) in Open Science II, a programme coordinated by swissuniversities. We are also grateful for DIZH-Support of the DSI PostDoc Project AI-R. We sincerely thank all participants of the UZH Shared Task, the organizers of the 13th Workshop on Argument Mining and Reasoning, especially Dr. Julia Romberg, for their valuable support and contribution to the successful organization of the UZH Shared Task.

References

Steven Bernstein. 2011. Legitimacy in Intergovernmental and Non-state Global Governance. *Review of International Political Economy : RIPE*, 18(1):17–51. Place: ABINGDON Publisher: Taylor & Francis Group tex.copyright: Copyright Taylor & Francis Group, LLC 2011.

Siddharth Bhargava. 2026. HybridArguer at UZH Shared Task 2026: Argument Structure Modeling in Bilingual UN Resolutions with Retrieval-Augmented and Iterative LLM Reasoning. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning, Budapest, Hungary. Association for Computational Linguistics.*

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.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

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.

Google. 2026. Gemma 4: New Open Models for Developers. Accessed: 2026-04-16.

Bogdan Grecu, Gerrit Quaremba, Elizabeth Black, Denny Vrandei, Elena Simperl, and Oana Cocarascu. 2026. Argchestrators at UZH Shared Task 2026: Efficient Argument Mining in UN Resolutions: A Sub-8B Pipeline using Agentic Debate and Heuristic Retrieval. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning, Budapest, Hungary. Association for Computational Linguistics.*

Vedant Gupta, Rahul Bhatia, Vaibhav Varshney, and Manjunatha Naik MC. 2026. RESOLVENOW at UZH Shared Task 2026: Rule-Based Type Classification with LLM-Driven Multi-Label Tagging for UN Resolutions. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning, Budapest, Hungary. Association for Computational Linguistics.*

Phuong Huu Vu Tran, Long Vo Minh, Son Nguyen Minh Le, and Hoang Van. 2026. LLM-INSTRUCT at UZH Shared Task 2026: Constraint-Aware Retrieval and Selective Debate for Paragraph-Level Argument Mining. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning, Budapest, Hungary. Association for Computational Linguistics.*

Rositsa V Ivanova and Reto Gubelmann. The Shift from Logic to Dialectic in Argumentation Theory: Implications for Computational Argument Quality Assessment. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4789–4802. Association for Computational Linguistics.

Siddhartha Khandelwal and Jyotsana Bhardwaj. 2026. Prompteam at UZH Shared Task 2026: RAG-Augmented Classification and Cosine-Filtered Re-

*https://diced.linguistik.uzh.ch/

lation Prediction for UN Resolutions. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning, Budapest, Hungary. Association for Computational Linguistics.*

Jyoti Kumari, Vinay Babu Ulli, Chandan Kumar R S, and Vaibhav Singh. 2026. TypeCoT at UZH Shared Task 2026: Reconstructing Argumentative Structure in UN Resolutions using Type-Informed Chain-of-Thought. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning, Budapest, Hungary. Association for Computational Linguistics.*

Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Kai Eckert. 2018. An Argument-Annotated Corpus of Scientific Publications. In *Proceedings of the 5th Workshop on Argument Mining*, pages 40–46, Brussels, Belgium. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. 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. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Patrick Montjouridès. 2022. Is this the Future We Want? Understanding the Legitimacy of International Education Agendas. The Example of Equity in Education. In *PhD Thesis*. University of Cambridge.

Margarete Sachs-Israel. 2016. The SDG 4-Education 2030 Agenda and Its Framework for Action - the Process of Its Development and First Steps in Taking It Forward. *Bildung und Erziehung*, 69(3):269–290. Place: Göttingen Publisher: Vandenhoeck und Ruprecht.

Giuseppina Scotto di Carlo. 2013. The Language of the UN: Vagueness in Security Council Resolutions Relating to the Second Gulf War. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 26(3):693–706.

Giuseppina Scotto di Carlo. 2017. Linguistic Patterns of Modality in UN Resolutions: The Role of Shall, Should, and May in Security Council Resolutions Relating to the Second Gulf War. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 30(2):223–244.

Sohom Sen, Avina Nakarmi, Xun Song, and Aritra Dasgupta. 2026. POINTERS at UZH Shared Task 2026: Evident: Reasoning Probes for Argumentation Mining in UN Resolutions. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning, Budapest, Hungary. Association for Computational Linguistics.*

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT–Building Open Translation Services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

UNESCO. 2012. *International Standard Classification of Education (ISCED) 2011*. UNESCO.

United Nations. 1983. United Nations Editorial Manual.

United Nations. 2025. United Nations Editorial Manual. Resolutions and other formal decisions of United Nations organs.

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, Valencia, Spain. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

A Human Gold Labels

Label	Round 1	Round 2
Modifying	135	166 (55%)
Complemental	111	79 (26%)
None	36	39 (13%)
Supporting	18	16 (5%)
Contradictive	0	0 (0%)
Total	300	300

Table 6: Human gold relation label distribution across annotation rounds. Round 2 is the authoritative gold; 233 of 300 pairs were blind re-annotated in round 2.

Team	Type mF1	Tag μF1	Tag mF1	Pair P	Pair R	Pair F1	Rel F1
Argchestrators	**0.901**	0.334	0.274	0.929	0.648	0.763	**0.380**
HybridArguer	0.857	0.390	0.251	0.901	0.628	0.740	0.278
LLM-Instruct	0.809	0.369	0.276	0.934	0.655	0.770	0.255
LLM-Instruct-2	0.856	0.317	0.240	0.915	0.207	0.338	0.372
Ockham	0.422	0.198	0.056	0.904	0.471	0.620	0.213
POINTERS	0.764	**0.493**	**0.413**	**0.947**	**0.751**	**0.838**	0.164
Prompteam	0.622	0.250	0.182	0.907	0.563	0.695	0.272
ResolveNow	0.877	0.353	0.240	0.000	0.000	0.000	—
TypeCoT	0.891	0.296	0.289	0.835	0.387	0.529	0.188

Table 7: Results against the human-annotated validation samples. μF1 denotes micro-averaged F1; mF1 denotes macro-averaged F1. Pair P/R/F1 measure pair identification within the 300-pair annotated sample; Rel F1 is weighted F1 over relation labels on correctly identified pairs only.

Complemental vs. Modifying: Disagreement Examples

The following pairs illustrate the boundary between *complemental* and *modifying* that accounts for 28% of round-2 label changes.

Resolution: ICPE-11-1948_RES1-FR_res_23

Round 1: *complemental* → Round 2: *modifying*

Para 12. A concrete, sensory, and motor initiation, offering the child numerous opportunities for creative activities, should **precede** for a sufficiently long period the acquisition of characters and the proper technique of writing;

Para 14. The learning of writing should take place **simultaneously** with that of reading, so that it has a lively and functional character;

Resolution: ICPE-06-1937_RES1-FR_res_10,

Round 1: *modifying* → Round 2: *complemental*

Para 17. 6) That, moreover, through organized trips abroad, internships and special courses, and by participating in the work of educational study commissions, in collaboration with professors from pedagogical institutes and normal schools, they may keep themselves informed of the developments in modern pedagogy;

Para 18. 7) That conferences enable them to establish among colleagues a certain unity of views compatible with the freedom of action of each of them;

B LLM-as-a-Judge Prompt

System prompt

You are an expert in argument mining, argument quality assessment, and philosophical logic. You will be given the thinking chains of LLMs that are engaged in an argument mining task. Your job is to use existing argument quality assessment frameworks to conduct an unbiased, grounded assessment of the argumentative quality of these thinking chains. You only rely on established metrics for argument quality assessment, you do not hallucinate and only focus on what is actually in the thinking chain.

How to proceed

1. Rely on established argument quality assessment metrics, especially Wachsmuth et al. (2017). Judge objectively what is present.

2. Use the following criteria:

(i) Logical Quality: Cogency or logical strength.
(ii) Rhetorical Quality: Persuasiveness, clarity, conciseness.
(iii) Dialectical Quality: Ability to resolve the issue for an informed audience.

3. Provide a short assessment for each criterion and assign a score (1–100). Then compute the average.

Grounding rules

Focus only on the thinking chains. Be direct, critical, and factually grounded. Use only the Wachsmuth et al. framework.

Output format

LOGICAL QUALITY (1–100)
[Assessment] Score: [1–100]

RHETORICAL QUALITY (1–100)
[Assessment] Score: [1–100]

DIALECTICAL QUALITY (1–100)
[Assessment] Score: [1–100]

FINAL SCORE
Average: [average]

Then provide scores as JSON.

Figure 3: LLM-as-a-judge evaluation prompt.