RESOLVENOW at UZH Shared Task 2026: Rule-Based Type Classification with LLM-Driven Multi-Label Tagging for UN Resolutions

Vedant Gupta[1] **Rahul Bhatia**[2] **Vaibhav Varshney**[2] **Manjunatha Naik**[2]
[1]Indian Institute of Technology Hyderabad [2]ServiceNow
ce23btech11059@iith.ac.in
{rahul.bhatia, vaibhav.varshney, manjunathanaik.mc}@servicenow.com

Abstract

Subtask 1 of the UZH Shared Task 2026 asks for paragraph-level classification of UN resolutions as preambular or operative and multi-label tagging from a 141-code, 15-dimension taxonomy, scored by tag F1 and an open-weight LLM-as-Judge on reasoning quality. Two earlier pipelines we built failed in opposite ways. An embedding-retrieval system dropped relevant tags before the LLM saw them. A per-dimension prompting system was accurate but too slow to iterate. The submitted system fixes both. A deterministic French-English lexical classifier assigns paragraph types at type macro-F1 of 0.910 on the official silver standard with no LLM calls. DeepSeek-R1-0528-Qwen3-8B (DeepSeek-AI, 2025) predicts tags through a single merged prompt that exposes the full taxonomy, hints likely dimensions, and walks the model through an eight-point checklist. ResolveNow places 7th overall, with 2nd on LLM-as-Judge and the 9th-place F1 rank explained by the absence of any Subtask 2 submission rather than by tagging quality on Subtask 1 in isolation. Code: https://github.com/wiesard-g12/shared-task.

1 Introduction

UN resolutions follow a near-deterministic surface structure. Preambular paragraphs open with present participles such as *rappelant* or *considérant*. Operative paragraphs open with indicative verbs such as *décide* or *recommande*. Type classification is close to a lookup. The hard half of Subtask 1 is the second one. Each paragraph must receive a subset of 141 multi-label tags across 15 thematic dimensions, and the correct tags often rest on indirect reference rather than lexical evidence. The shared task scores submissions by averaging tag F1 with an LLM-as-Judge on free-text reasoning quality, and restricts systems to open-weight models at most 8B parameters.

Two natural designs fail. Narrowing 141 candidate tags by embedding retrieval throws away recall on a code space too small for retrieval to add value. Issuing one LLM call per dimension preserves coverage but multiplies inference cost by 5x to 6x, blocking iteration. The submitted system resolves both failures. A merged prompt exposes the full taxonomy in a single call, an eight-point checklist forces dimension-by-dimension coverage, paragraphs are batched three per call with isolation fencing, and rule-based type classification frees the LLM budget for tags. The system places 7th overall, with 2nd on LLM-as-Judge and 9th on the combined F1 axis. The combined F1 ranking aggregates Subtask 1 tag F1 with Subtask 2 relation-prediction F1, and ResolveNow submitted only to Subtask 1, so the 9th-place combined F1 rank reflects that absence rather than tagging quality.

2 Task and Data

The training set is 2,695 UN resolutions in French with Opus-MT English translations, drawn from the UN archive corpus introduced for prior work on religion and spirituality retrieval (Gao et al., 2025). The held-out test set is 45 annotated UNESCO International Bureau of Education resolutions (1934–2008), roughly 1,350 paragraphs in total, with English translations. Each paragraph carries three target fields: a `type` label, a list of `tags` from the 141-code taxonomy, and a free-text `think` field scored 0–100 by an open-weight LLM-as-Judge. The taxonomy CSV covers education level, teachers, policy theme, learner population, curriculum, pedagogy and assessment, infrastructure, legal frameworks, cross-cutting themes, system monitoring, learning modality, vocational vs. general orientation, ownership, stakeholders, and subject domains. A paragraph receives tags from several dimensions at once. Population tags are the most restrictive axis, requiring explicit nam-

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 125–128

ing of the group rather than implicit reference.

3 System

3.1 Pipeline

Stage 1 is a deterministic French-English lexical classifier that assigns paragraph type from the opening words. The lexicon contains over 100 French and 60 English patterns covering present-participle preambular openers, indicative-verb operative openers, and conference-style multi-word formulas. A three-step cascade strips numeric prefixes, matches the longest French phrase among the first five words, then falls back to English. Paragraphs that match no lexical pattern are resolved by position relative to the first confidently operative paragraph. The classifier issues no LLM calls.

In Stage 2, the paragraph, its type label from Stage 1, and the full 141-code taxonomy pass to DeepSeek-R1-0528-Qwen3-8B (DeepSeek-AI, 2025), an 8B reasoning-tuned distillation of DeepSeek-R1 into a Qwen3-family architecture (Yang et al., 2024). The model returns a JSON object with selected tag codes and a structured per-paragraph `reasoning` field. Its native `<think>` traces serve double duty. They sharpen tag selection and populate the field the LLM-as-Judge evaluates.

3.2 Merged Prompt

The prompt presents the full taxonomy with all 15 dimensions, every code, and a one-line semantic description per code. An eight-point checklist orders the reasoning across dimensions: stakeholders and populations, education levels, teachers, policy themes, curriculum and infrastructure, cross-cutting themes, modality and ownership, monitoring. Without the checklist the model tags whichever dimension is lexically obvious and skips the rest. With it, dimension coverage improved markedly during development.

Roughly 60 regex patterns scan each paragraph in both languages and flag likely dimensions as a single advisory hint line. Hints are not restrictive. The full taxonomy stays visible and the checklist forces systematic coverage. Five disambiguation rules address tag categories the model confuses: `POP_*` requires explicit mention of the group, `ACT_EDUC` refers to teachers as educators rather than school administrators (those fall under `ACT_GOV`), `ISC_*` requires explicit level mention, `O_VET` refers only to vocational programs, and `M_NFORM` refers to post-school education. The prompt biases toward recall, since missing a valid tag is penalized more than including a borderline one.

3.3 Batching and Inference

Three paragraphs are batched per LLM call, cutting calls from roughly 1,350 to 450 on the test set. The risk is contamination across paragraphs in the same context window. Three mechanisms contain it. Each paragraph is wrapped in a visually distinct fence with an explicit index. The prompt instructs the model to treat each paragraph as if from a separate document. The output schema requires one JSON object with an independent `reasoning` field per paragraph, removing the structural option to emit a blended answer. When a batch returns malformed JSON or omits a paragraph, the system retries the missing item as a single call. Inference runs in bfloat16 on a single A100 80GB GPU.

3.4 Pipeline Evolution

Table 1 summarizes the three architectures we tried.

4 Results

The split between the two evaluation axes is the central finding. A 2nd-place finish on LLM-as-Judge alongside a 9th-place combined F1 finish (Table 2) places the reasoning traces among the strongest in the competition while leaving the system's combined F1 ranking last. The combined F1 axis aggregates Subtask 1 tag F1 with Subtask 2 relation-prediction F1, with Subtask 2 relation-prediction carrying roughly 60% of that ranking. ResolveNow submitted only to Subtask 1, and on Subtask 1 tag metrics in isolation the system places 5th. The 9th-place combined F1 ranking therefore reflects the absence of any Subtask 2 submission rather than a failure of the tagging pipeline.

A structured per-paragraph `reasoning` field walks through each of the 15 dimensions and justifies each decision, giving the judge a consistent rationale to evaluate. This explains the 2nd-place finish on LLM-as-Judge. Tag selection runs in a single zero-shot pass without retrieval or training-data exemplars, which limits coverage on paragraphs where the relevant tags rest on indirect evidence. The `POP_*` disambiguation rule trades recall for precision on implicit population mentions, which contributes to the residual tag F1 gap relative to the top systems on Subtask 1 alone.

Method	Architecture	Reason for change
Method 1	Embedding retrieval narrows the 141 tags to a candidate shortlist, then a batched LLM call selects from the shortlist. Each tag code in the CSV was first enriched with a one-sentence description of what the code denotes (e.g., `ISC_01` as "care and stimulation programs for infants and toddlers before they reach preschool age") so paragraph embeddings were matched against meaning rather than opaque codes.	Even with enriched descriptions, retrieval dropped relevant tags before the LLM saw them. The 141-code space is too small for retrieval to improve precision without hurting recall.
Method 2	Per-paragraph two-stage prompting. Stage 1 picks the relevant dimensions out of 15. Stage 2 issues one LLM call per selected dimension and chooses tags within it.	Tagging accuracy was acceptable but the dimension expansion was too slow for prompt iteration on a single A100, blocking refinement of disambiguation rules.
Method 3	Single merged prompt with the full taxonomy, keyword hints, and an eight-point checklist. Three paragraphs batched per call with isolation fencing. Type classification moved out of the LLM into rules.	The merged prompt recovers coverage across all 15 dimensions in a single pass. Batching and rule-based types bring inference back within practical bounds. Submitted system.

Table 1: Iterations of the tagging pipeline. Method 1 showed that narrowing the candidate set is risky for a 141-code space even with semantic enrichment. Method 2 showed that per-dimension prompting scales poorly. Method 3 keeps Method 2's coverage guarantee through checklist scaffolding, avoids Method 1's retrieval risk by exposing the full taxonomy, and recovers speed by batching and offloading types to rules.

Team	F1	Judge	Final
LLM-Instruct	1	5	1
Prompteam	5	1	2
Argchestrators	2	6	3
HybridArguer	4	3	3
POINTERS	3	9	5
ResolveNow	**9**	**2**	**7**
TypeCoT	6	8	8
Ockham	8	7	9

Table 2: Official leaderboard. Columns are per-metric ranks. F1 aggregates Subtask 1 tag F1 with Subtask 2 relation-prediction F1.

On paragraph type, the rule-based classifier reaches type macro-F1 of 0.910 on the official silver standard. Residual errors come from rare participial variants outside the lexicon (handled by the positional fallback, occasionally mislabeled at document boundaries) and from documents written entirely as numbered lists without verb-initial openers (handled by a Format-B fallback that defaults to operative).

5 Related Work

Argument mining on political and legal text has focused on claim-premise extraction and stance, with little attention to paragraph-level role labeling in UN-style resolutions. The corpus released alongside SpiritRAG (Gao et al., 2025) provides large-scale annotated UN archive material and underlies the training data used in this shared task. The shared task formulation pairs structural labels with a 141-code topical taxonomy across 15 dimensions, reframing the problem as both structural and multi-label topical. Multi-label classification over large taxonomies has shifted from supervised classifiers toward prompted LLM inference, both for zero-shot reach and for compatibility with reasoning-quality evaluation. Reasoning-distilled models such as DeepSeek-R1 (DeepSeek-AI, 2025) expose intermediate reasoning traces directly, which the LLM-as-Judge metric rewards. Our merged prompt with explicit checklist scaffolding extends the structured-prompting pattern to a 15-dimension, 141-code taxonomy in a single inference pass.

6 Conclusion

Splitting Subtask 1 along the rule-versus-reasoning boundary, then exposing the full 141-code taxonomy to an 8B reasoning model with explicit dimension scaffolding, achieves competitive reasoning quality (2nd on LLM-as-Judge) at substantially lower inference cost than per-dimension prompting. The 9th-place combined F1 ranking traces to the absence of a Subtask 2 submission rather than to the tagging pipeline itself, which places 5th on

Subtask 1 tag metrics in isolation. Closing the combined F1 gap requires extending the system to the relation-prediction subtask. The residual tag F1 gap on Subtask 1 points toward retrieval-augmented or few-shot grounding as the natural next direction.

Limitations

We submitted only to Subtask 1. The relation-prediction Subtask 2 would require a separate pipeline and was not addressed. The rule-based type classifier works because UN French drafting is formulaic, and would need a new lexicon for any other organization or drafting tradition. The 2,695 training resolutions were used neither for fine-tuning nor for few-shot retrieval. The reasoning scaffold is working, and only the tag selection given that scaffold needs help on Subtask 1 in isolation, which points directly at retrieval-augmented generation as the missing component there. Batch-3 grouping by document order places thematically related paragraphs in the same call, which is the worst case for isolation fencing. Grouping by content dissimilarity would mitigate this. The 8B parameter cap limits achievable reasoning quality, and scaling to larger open models is expected to raise tag F1 at proportional inference cost.

References

Yingqiang Gao, Fabian Winiger, Patrick Montjouridès, Anastassia Shaitarova, Nianlong Gu, Simon Peng-Keller, and Gerold Schneider. 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, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645:633–638, 2025. https://doi.org/10.1038/s41586-025-09422-z.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024.