ArgMining 2026

13th Workshop on Argument Mining and Reasoning

Proceedings of the 13th Workshop on Argument Mining and Reasoning

July 3, 2026

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-399-9

Introduction

Argument mining (also known as “argumentation mining”) is a well-established research area in computational linguistics that focuses on the automatic identification of argumentative structures, such as premises, conclusions, and inference schemes. Since its beginnings, the focus has been on the development of large-scale argumentation datasets and tasks like argument quality assessment, argument persuasiveness, and the synthesis of argumentative texts, spanning various domains, such as legal, social, medical, political, and scientific settings.

Mirroring the advancements in CL and NLP at large, argument mining has started to work on explainable argumentation, multimodal settings, and modeling human label variation. The field has also started to investigate the performance of generative models in producing and analyzing human-like argumentation. The quality of those models depends heavily on the task; for instance, LLMs show strong performance in generating persuasive essays and convincing arguments, but they have been shown to fall short in identifying fallacious arguments and to reproduce biases inherent in models, requiring substantial fine-tuning efforts to mitigate them. It remains largely unexplored whether LLMs truly possess a deeper understanding of argumentation, or whether they are simply effective pattern learners.

The 2026 edition of the ArgMining workshop therefore places a **special focus on understanding and evaluating arguments in both human and machine reasoning**. With this topic, we broaden the workshop’s focus to include reasoning, a long-standing area of research in AI that has recently gained renewed interest within the *ACL community, driven by the latest generation of LLMs. Reasoning is tightly connected to argumentation as it represents, analyzes and evaluates the process of reaching conclusions on the basis of available information. If we consider argumentation as a paradigm to capture reasoning, then machines (particularly LLMs) can be evaluated based on their ability to address argument mining tasks.

ArgMining 2026 received 20 submissions in the main workshop track, 9 non-archival submissions and 8 submissions to the Shared Task track. The final program of the workshop consists of 7 oral main workshop papers, 2 short papers as posters, 6 non-archival papers as posters and 7 Shared Task posters.

Addressing this year’s topic of understanding and evaluating arguments in both human and machine reasoning, Lora Aroyo from Google Research in NYC gave the invited talk. The workshop also featured a panel on **Argument Mining Meets Reasoning: Understanding and Evaluating Arguments in Both Human and Machine Reasoning.**

Our thanks go to everyone who contributed to realizing this workshop, in particular the invited speaker, the panelists, the authors and the Programme Committee.

Mohamed Elaraby, Annette Hautli-Janisz, John Lawrence, Elena Musi, Julia Romberg, Federico Ruggeri
(ArgMining 2026 co-chairs)

Organizing Committee

Workshop Organizers

Mohamed Elaraby, University of Pittsburgh
Annette Hautli-Janisz, University of Passau
John Lawrence, University of Dundee
Elena Musi, University of Liverpool
Julia Romberg, GESIS Leibniz Institute for the Social Sciences
Federico Ruggeri, University of Bologna

Shared Task Organizers

Yingqiang Gao, University of Zurich, Switzerland
Anastassia Shaitarova, University of Zurich, Switzerland
Reto Gubelmann, University of Zurich, Switzerland
Patrick Montjouridès, University of Zurich, Switzerland

Program Committee

Program Committee

Aldo Costa, Heinrich-Heine Universität Düsseldorf
Ashish Anand, Indian Institute of Technology, Guwahati
Benno Stein, Bauhaus Universität Weimar
Blanca Calvo Figueras, Universidad del País Vasco
Davide Liga, University of Luxemburg
Deborah Dore, Université de Nice-Sophia Antipolis
Dexter Williams, University of Wisconsin - Madison
Eimear Maguire, University of Dundee
Elena Cabrio, Université Côte d'Azur
Elena Chistova, FRC CSC RAS
Eva Maria Vecchi, University of Stuttgart, Universität Stuttgart
Giulia D'Agostino, Fraunhofer FKIE
Giulia Grundler, University of Bologna
Henning Wachsmuth, Leibniz Universität Hannover
Irina Nikishina, Universität Hamburg
Johannes Kiesel, GESIS – Leibniz Institute for the Social Sciences
Joonsuk Park, University of Richmond
Khalid Al Khatib, University of Groningen
Lynn Greschner, Otto-Friedrich Universität Bamberg
Manfred Stede, Universität Potsdam
Marc Feger, Heinrich-Heine Universität Düsseldorf
Marco Lippi, University of Florence
Matthias Orlikowski, Universität Bielefeld
Maximilian Martin Maurer, GESIS Leibniz Institute for the Social Sciences
Mohammad Sharabati, Heinrich-Heine Universität Düsseldorf
Neele Falk, University of Stuttgart, Universität Stuttgart
Patrick Saint-Dizier, CNRS
Rabiraj Bandyopadhyay, GESIS – Leibniz Institute for the Social Sciences
Ramon Ruiz-Dolz, University of Dundee
Tangming Yuan, University of York
Yamen Ajjour, University of Technology Nuremberg
Zlata Kikteva, Universität Passau

Invited Speaker

Lora Aroyo, Google Research, United States

Table of Contents

Program

Friday, July 3, 2026

08:45 - 09:00 *Opening Remarks*

09:00 - 10:30 *Session I*

A Three-Level Audit of LLM Alignment for Argument Quality Assessment
Wei-Fan Chen, Jinming Yu and Lucie Flek

TypeCoT at UZH Shared Task 2026: Reconstructing Argumentative Structure in UN Resolutions using Type-Informed Chain-of-Thought
Chandan Kumar R S, Vinay Babu Ulli, Jyoti Kumari and Vaibhav Singh

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

Illustrating Arguments with Images Using Aspect-Aware Prompting
Maximilian Heinrich, Sharat Anand, Johannes Kiesel and Benno Stein

A Neural Approach to Fine-Grained Argumentation Strategy Classification with Emotion and Moral Value Lexicons across Multiple Domains
Mohammad Yeghaneh Abkenar, Weixing Wang, Manfred Stede and Julia Romberg

10:30 - 11:30 *Coffee Break*

11:30 - 12:30 *Keynote: Lora Aroyo*

12:30 - 14:00 *Lunch Break*

14:00 - 14:30 *Shared Task Overview, Best System, Awards*

14:30 - 15:30 *Panel Discussion: Argument Mining Meets Reasoning: Understanding and Evaluating Arguments in Both Human and Machine Reasoning.*

15:30 - 16:30 *Coffee Break*

15:30 - 17:00 *Poster Session (Main Workshop Papers, Non-Archival Papers, and Shared Task Papers)*

17:00 - 17:30 *Session II*

RESOLVENOW at UZH Shared Task 2026: Rule-Based Type Classification with LLM-Driven Multi-Label Tagging for UN Resolutions
Vedant Gupta, Rahul Bhatia, Vaibhav Varshney and Manjunatha Naik

Beyond Logical Forms: LLM-Extracted Patterns for Fallacy Classification
Eleni Papadopulos, Firoj Alam and Giovanni Da San Martino

17:30 - 17:45 *Closing Remarks*

STCOR: A Trilevel Syllogism-Driven Reasoning Framework

Keying Yang[1,2], **Hao Wang**[1], **Chengtao Jian**[1], **Kai Yang**[1,2,3†]
[1]School of Computer Science and Technology, Tongji University, China.
[2]Shenzhen Loop Area Institute, China.
[3]Shanghai Research Institute for Intelligent Autonomous Systems, China
{yky, 2432196, jct, kaiyang}@tongji.edu.cn
† Corresponding author.

Abstract

Inspired by the human expert thinking paradigm in operations research, this work introduces a new concept of reasoning tasks: Textual Constrained Optimization (TCO) problems. A TCO problem is characterized by a natural language description that implicitly specifies an underlying structured model with variables, constraints, and objectives. Then, we propose a novel Syllogism-driven Textual Constrained Optimization Reasoning (STCOR) paradigm, which is driven by classical syllogistic logic. Unlike contemporary stepwise methods, our framework structures reasoning into three phases: meta-modeling, which acts as the major premise by retrieving a relevant class-driven prototype template; formalization, which serves as the minor premise by instantiating the template into an explicit logical model from textual queries; and solving, which derives the final answer as conclusion. To support the end to end implementation, we further develop a tri-level optimization algorithm TriRL. This algorithm enables the joint training of all core components to ensure the coherence and efficiency of the entire reasoning system. Extensive experiments on multiple benchmarks, including a TCO benchmark we developed, demonstrate that STCOR achieves significant accuracy improvements (over 10% on average) while enabling the explicit tracing and rectification of reasoning leaks and logical fallacies.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive advances in reasoning tasks, significantly attributed to Chain-of-Thought (CoT) (Wei et al., 2022). Some advanced reasoning models such as R1 (Guo et al., 2025) and O1 (Jaech et al., 2024), further improve accuracy by extending CoT length. However, these methods predominantly rely on surface-level pattern recognition, lacking mechanisms to capture the problem's underlying logical structure, which limitation has fueled critiques of LLMs as "stochastic parrots" (Stechly et al., 2023) and the inability to identify reasoning fallacies as the inference chain grows. Moreover, in real-world applications like law and finance, users require not only correct answers, but also clear, auditable, and traceable reasoning processes to assess trustworthiness (Roychowdhury, 2024).

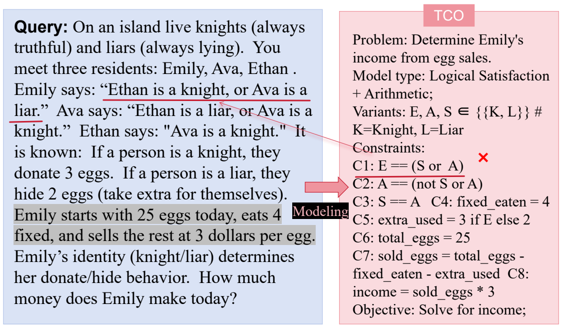


Figure 1: A sample problem from our TCO-Bench (Easy), which blends logical and arithmetic constraints. The greyed-out text shows the original GSM8K problem; we transform it into a hybrid puzzle requiring both logical deduction and numeric calculation. By explicitly reconstructing the underlying TCO model, it provides a transparent, step-wise reasoning trace where each constraint can be verified and errors can be precisely localized.

Unlike step-by-step reasoning, human experts typically begin by abstracting problems into formal models (e.g., variables, constraints, objectives) grounded in first principles and then seek solutions by applying analytical or computational methods. This *"modeling and solving"* paradigm, prevalent in mathematics and operations research areas, enables generalizable and traceable human reasoning.

Although code-augmented reasoning leverages computational precision, existing methods, ranging from general-purpose Python (Chen et al., 2022; Gao et al., 2023; Gou et al., 2024) to logic programming (Pan et al., 2023; Ye et al., 2023; Zhou et al., 2024b), predominantly rely on a direct translation of reasoning steps into programs. These methods

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 1–11
July 3, 2026 ©2026 Association for Computational Linguistics

often fail to disentangle implicit reasoning signals from noisy, abstract inputs, confining its success to narrow domains like mathematics. Furthermore, due to the imbalanced distribution of pre-training data, the entanglement between reasoning logic and code syntax gives rise to performance variance across different programming languages (Xu et al., 2022; Li et al., 2025).

Therefore, inspired by this expert paradigm, we identify and formally define a critical class of Textual Constrained Optimization (TCO) problems, which are defined by a natural language problem that implicitly specifies a formal model with *variables, constraints, and objectives*. A TCO problem typically involves constraints expressed through rules, logic, semantics, or mathematical relations, along with an objective that entails decision-making, judgment, or the derivation of a unique solution. This definition provides a new perspective on interpreting argument mining of original problem. To better capture its characteristics, we construct a TCO benchmark by rewriting and extending problems from GSM8K (Cobbe et al., 2021).

Moreover, we introduce a novel reasoning paradigm: **S**yllogism-driven **T**extual **C**onstrained **O**ptimization **R**easoning (**STCOR**), which is inspired by the classical concept of generalized syllogism. STCOR decomposes the reasoning process into a principled, three-level argumentative reconstruction layer: (i) A meta-modeling layer establishes the "Major Premise" by selecting an optimal class-driven prototype template from libraries. (ii) A formalization layer constructs the "Minor Premise" by transforming natural language queries into structured, formal TCO models. (iii) A solving layer derives the "Conclusion" by generating and executing deterministic code. This framework forms a closed syllogistic loop, ensuring reasoning is traceable, auditable, and grounded in logic. To enhance meta-modeling, we design a prototype knowledge library, composed of class-driven prototype templates to capture common categories and structural patterns of TCO problems.

Furthermore, to enhance generalization, we design a tri-level end-to-end optimization algorithm TriRL, based on group relative policy optimization (GRPO) (Shao et al., 2024) algorithm. TriRL enables joint training of STCOR's three core components, ensuring systemic coherence and efficiency. Extensive experiments on 9 datasets (math, causal, logical, commonsense reasoning datasets and our benchmark) with multiple backbone LLMs show that STCOR achieves an average absolute accuracy gain of over 10% compared to the strongest baselines.

Our contributions are summarized as follows:

- We formally identify and define a new class of problems, termed Textual Constrained Optimization (TCO). To better capture the features, we developed a TCO benchmark dataset.

- We propose a novel Syllogism-driven TCO Reasoning (STCOR) paradigm and instantiate it with a tri-level framework that decomposes reasoning into meta-modeling, formalization, and solving. Also, we construct a class-driven prototype templates library for meta-modeling.

- We design TriRL, a tri-level optimization algorithm for the end-to-end joint training of STCOR framework. Extensive experiments validate STCOR's superior accuracy and interpretability, showcasing its effectiveness on different tasks.

2 Related Work

LLM Reasoning Paradigm Starting from the Chain-of-Thought (CoT) (Wei et al., 2022), which proposes to elicit LLM thinking step by step, different reasoning paradigms have emerged to bolster reasoning performance, mainly including: planning (Hao et al., 2023; Wang et al., 2023), problem decomposition (Zhou et al., 2024a), analogical reasoning (Yang et al., 2024b; Yu et al., 2024), self-improvement (Tian et al., 2024), abstraction (Hong et al., 2024), searching (Besta et al., 2024; Yao et al., 2023), inference scaling (Snell et al., 2024), meta-thinking (Xiang et al., 2025; Wan et al., 2025) and programming (Gao et al., 2023; Chen et al., 2022). Similar to meta-thinking or reflection, modeling is also a type of advanced human cognition. In this work, we propose a novel tri-level framework with the aim of stimulating the abstract modeling ability of LLMs.

Code-Enhanced Reasoning Code's computational and logical strengths have made it a powerful approach for enhancing LLM reasoning. Some works generate executable code through fine-tuning or prompting, mainly including general-purpose languages such as Python (Chen et al., 2022; Gao et al., 2023; Gou et al., 2024) and logic programming languages (Pan et al., 2023; Ye et al., 2023; Zhou et al., 2024b). As a specialized category of mathematical problems, recent studies have focused on employing multi-agent frameworks to address operations research problems (Wang et al., 2025; Xie et al.; Xiao et al., 2023). However, these works remain confined to narrow domains (e.g., math, symbolic tasks) due to the challenges in disentangling implicit reasoning signals from noisy, ambiguous information for abstract, complex input problems. Besides, due to the imbalanced distribution of pre-training data, the entanglement between reasoning logic and code syntax gives rise to performance variance across different programming languages (Xu et al., 2022; Li et al., 2025). Our framework unleashes the power of code-augmented reasoning across a broader range of tasks.

Reinforcement Learning for LLMs To improve the performance, RLHF methods like Proximal Policy Optimization (PPO) (Schulman et al., 2017; Ouyang et al., 2022) are introduced for LLM alignment. Then direct Preference Optimization (DPO) (Rafailov et al., 2023) and its variants were (Jian et al., 2025) was proposed to simplify by directly optimizing with pairwise data. Recent techniques like Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and REINFORCE++ (Hu, 2025) aim to reduce the computational cost of critic networks in PPO. However, long contexts and extended inference trajectories remain impractical for smaller models. In this paper, we design a tri-level reinforcement learning optimization algorithm to jointly optimize our framework.

3 Preliminaries

Just as logistics optimization must be translated into formal operations research (OR) models for systematic solving, textual reasoning problems often conceal a structured model whose core elements are obscured by natural language. This gap motivates us to formalize a specific category of problems termed **Textual Constrained Optimization (TCO)**.

3.1 Textual Constrained Optimization Problem

Definition 1 (**Textual Constrained Optimization (TCO) Problem**). *A TCO problem q is a natural language problem whose solution requires the discovery of a latent, structured model with five-tuple:*

$$\mathcal{M} = (p, t, \mathcal{V}, \mathcal{C}, \mathcal{O}) \quad (1)$$

where:

- *p ("problem overview") is a brief summary of the task,*
- *t ("model type") indicates the model category (e.g., causal inference, SAT, CSP);*
- *$\mathcal{V} = \{v_1, \dots, v_n\}$ is a set of decision or state variables,*
- *$\mathcal{C} = \{c_1, \dots, c_m\}$ is a set of constraints (e.g., equations, inequalities, logical relations) over $\mathcal{V}$,*
- *O ("objective") specifies the goal(s) defined over $\mathcal{V}$, (e.g., "calculate x" or "determine if x=true").*

As shown in Figure 1, we adopt an *optimization-perspective syntax* to formally describe the model $\mathcal{M}$, allowing a mild mixture of textual and symbolic expressions instead of rigid mathematical formulas. Unlike formal languages with specific syntactic forms (e.g., ATP (Zhou et al., 2024b)), this method strikes a balance between formal representational capacity and flexibility, enabling stronger expressive power.

3.2 Syllogism-driven TCO (STCO) Reasoning

Given a TCO problem q, now the core challenge shifts to *how to reliably construct and solve the latent model $\mathcal{M}$*. Inspired by classical syllogistic logic and the expert "modeling-and-solving" pipeline, we propose the Syllogism-driven Textual Constrained Optimization Reasoning (STCOR) paradigm.

Definition 2 (**Syllogism-driven TCO Reasoning Paradigm**). *Given a TCO problem query $q \in \mathcal{Q}$, our goal is to derive the final answer $a \in \mathcal{A}$. The Syllogism-driven TCO Reasoning Paradigm structures the solution process for a TCO problem q into three syllogism-inspired phases:*

$$\mathcal{Q} \xrightarrow{\alpha} \mathcal{T}^* \xrightarrow{\beta} \mathcal{M} \xrightarrow{\gamma} (\mathcal{P}, \mathcal{A}) \quad (2)$$

1. ***Meta-Modeling** (α): From the textual query q, retrieve a corresponding high-level prototype template T^* as "Major Premise".*

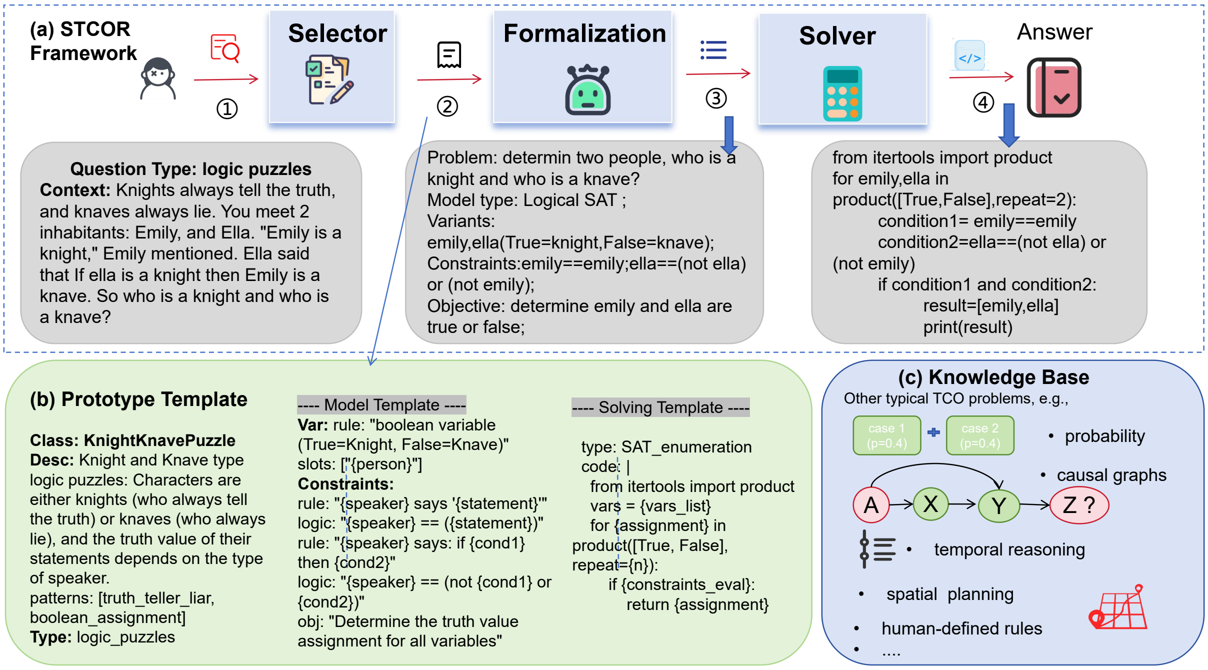


Figure 2: Overview of the STCOR reasoning pipeline and its core components.(a) Syllogistic three-stage reasoning: Meta-modeling retrieves a template; formalization instantiates it into an explicit logical model; solving generate solution code to produce the final answer. (b) Class-driven prototype template: An example template encoding the high-level abstract structural schema and solving strategy. (c) Prototype knowledge base: A collection of such templates across diverse TCO types.

2. ***Formalization*** *(β): Using the template T^* as a schema, ground the specifics of q into a formal model instance $\mathcal{M}$ as "Minor Premise".*

3. ***Solving*** *(γ): Derives the final answer a from model $\mathcal{M}$ via a deterministic solving process $p \in \mathcal{P}$ (e.g., generating and executing code), yielding the "Conclusion".*

This paradigm ensures that reasoning is grounded in first principles, providing a transparent and auditable chain from problem to solution. The following sections introduce the **STCOR framework**, a concrete computational instantiation of this paradigm, detailing the learnable components (see §4) and the optimization algorithm (see §5) that realize these phases.

4 STCOR Framework

As shown in Figure 2, the STCOR framework consists of three components: a selector that retrieves the most relevant prototype template T in meta-modeling layer; a formalization LLM that instantiates the structured formal model $\mathcal{M}$; and a solver LLM that generates and executes code guided by template T and problem $\mathcal{M}$. This decoupled design achieves modular specialization, interpretable intermediate representations, and reduced context complexity for each stage.

4.1 Meta-Modeling with Prototype Knowledge Base

Prototype Knowledge Base When solving practical OR problems, human experts typically do not start from scratch; instead, they rely on two key forms of prior knowledge: (1) identifying the problem as an instance of a canonical type (e.g., scheduling or inventory management), and (2) knowing the standard modeling and solution paradigms for that type (e.g., integer programming for scheduling). Inspired by this, we observe that a wide range of general reasoning tasks (e.g., rule-based reasoning, mathematical computation, logical deduction, causal inference) can be formulated as TCO problems. We categorize typical TCO problems by the dominant type of constraint into five core types:

- **Logical**: e.g., truth values, implication, equivalence.
- **Mathematical**: e.g., algebraic equations, inequalities.
- **Probabilistic**: e.g., probability calculations.

- **Rule-based**: e.g., causal graphs, human-defined rules.

- **Commonsense**: e.g., spatial, temporal commonsense.

A single TCO problem can involve a mixture of these constraints. To provide reusable *prior knowledge* for typical TCO problem archetypes, we construct a prototype knowledge base $\mathcal{K} = \{T_1, T_2, ..., T_K\}$. Each $T_i \in \mathcal{K}$ is a structured **prototype template** corresponding to a typical TCO problem (e.g., mathematical puzzles, logical riddles, causal networks, temporal reasoning, spatial planning and hybrid problems).

Class-driven Prototype Template Different from the thought templates in BoT (Yang et al., 2024b) that focus on reasoning steps, our prototype templates aim to capture the structural essence of a problem (e.g., attributes, logical relations, core patterns). A prototype refers to a canonical, representative example or pattern that summarizes the common characteristics of a class of entities. For example, like class-object concept in object-oriented programming: a class acts as the prototype (defining structure and behavior), and an object is its instance. To abstract general "major-premise" guiding principles, we propose a *class-driven prototype template*. Formally, a prototype template T_i can be represented as:

$$T_i = (\text{desc}_i,\ t_i,\ \tau_i.\ \sigma_i) \tag{3}$$

where desc_i denotes a brief description of the problem class's features, acting as a semantic anchor for retrieval and interpretation; t_i represents the semantic type identifier (e.g., logic_puzzles); τ_i refers to modeling knowledge, consisting of a set of constraint and objective patterns; and σ_i indicates solving knowledge, which specifies the recommended solving strategy.

Selector and Meta-modeling Given an input query q, the meta-modeling stage aims to retrieve the most relevant prototype template T^* from the knowledge base $\mathcal{K}$. Since a semantic gap between the abstract template descriptions and the narrative problem instances, we adopt a two-stage retrieval–ranking architecture. In retrieval stage, we construct a candidate template set $\mathbb{T}_{\text{can}} = \{T_i\}_{i=1}^{k}$ via a hybrid strategy to ensure both reliability (from retrieval) and creativity (from generation). First, templates are retrieved from the predefined library $\mathcal{K}$ based on semantic similarity:

$$\mathbb{T}_{\text{re}} = \text{Top-}k'_1\left(\{T_i \in \mathcal{K} \mid \text{Sim}(f(q), f(D_{T_i})) \geq \delta\}\right) \tag{4}$$

where $f(\cdot)$ is a text embedding model, D_{T_i} represents the description part of T_i, $\text{Sim}(\cdot)$ computes cosine similarity, and δ is the similarity threshold. The Top-$k'_1(\cdot)$ operation returns the top k'_1 templates with the highest similarity, and $k'_1 = \min(k_1,\ |\{T_i \mid \text{Sim}(\cdot) \geq \delta\}|)$. If $\mathbb{T}_{\text{re}} = \emptyset$, the current task q is identified as a new task. Subsequently, to supplement the candidate set size, a generative language model is guided to generalize potential solution rules for the current problem and generate $k_2 = k - k'_1$ candidate templates: $\mathbb{T}_{\text{gen}} = \text{LLM}_{\text{propose}}(q, k_2)$.

Then in the ranking stage, a lightweight model π_{θ_α} (e.g., BERT) selects the optimal template from the merged candidate set:

$$T^* = \arg\max_{T_i \in \mathbb{T}_{\text{can}}} \pi_{\theta_\alpha}(q, T_i) \tag{5}$$

where $\mathbb{T}_{\text{can}} = \mathbb{T}_{\text{re}} \cup \mathbb{T}_{\text{gen}}$.

4.2 Optimization-guided Formalization

Guided by the modeling rules of the meta-modeling template T^*, the formalization layer serves as the core engine for converting a textual problem q into a solvable TCO instance. This process is implemented by a policy π_{θ_β}, which performs structured reasoning to produce a formalization trace $\mathcal{S}$ and the resulting model $\mathcal{M}$ as defined in §3.1:

$$(\mathcal{S}, \mathcal{M}) = \pi_{\theta_\beta}(q, T^*) \tag{6}$$

where $\mathcal{S} = (s_1, s_2, \ldots, s_l)$ is a step-by-step rationale documenting how each element of $\mathcal{M}$ is derived from q and T^*. This process yields a formal model $\mathcal{M}$ that not only captures the essential problem structure but also provides an auditable trace directly back to the original problem textual.

4.3 Solving and Execution

Based on related solution strategy in T^*, this layer transforms the formal model $\mathcal{M}$ into an executable solution. Given a solver policy π_{θ_γ}, it generates a program p (e.g., in Python), which is executed by an external solver $\mathcal{L}$ to derive the final answer a. This process is formalized as:

$$p = \pi_{\theta_\gamma}(\mathcal{M}, T^*), \quad a = \mathcal{L}(p) \tag{7}$$

where $\mathcal{L}$ denotes the external tool.

5 Algorithm

To further strengthen the training of our framework, we propose an end-to-end tri-level algorithm under reinforcement learning settings.

5.1 Problem Formulation

First, we formalize this as a tri-level optimization problem (Huang et al., 2024; Jian et al., 2024; Jiao et al., 2024, 2025; Zhu and Yang, 2026): The meta-modeling selector π_{θ_α} is to select the optimal template T^* that maximizes the subsequent formalization and solving performance. Based on standard policy gradient, its objective is to maximize the expected reward, defined as:

$$\mathcal{J}_{\text{meta}}(\theta_\alpha) = \mathbb{E}_{T^* \sim \pi_{\theta_\alpha}(\cdot|q)} \left[r_T(T^*, q) \right] \tag{8}$$

where $r_T(\cdot)$ denotes the utility reward of template T^* for problem q. It is computed as the average reward of formal models generated under the guidance of T^*, i.e., $r_T(T^*, q) = \frac{1}{G}\sum_{i=1}^{G} r_m(m_i)$, where $\{m_i\} \in \mathcal{M}$ are models generated by the formalization layer, and $r_m(m_i)$ is the reward of model m_i.

Given a selected template T^* and a problem $q \sim P(Q)$, the formalization layer first samples G candidate models $\{m_i\}$ from $\pi_{\theta_\beta^{\text{old}}}(m|q)$, and the solving layer then samples P candidate outputs $\{o_{ij}\}$ from $\pi_{\theta_\gamma^{\text{old}}}(o|m_i)$. Considering the GRPO objective (Shao et al., 2024) (without KL penalties), its objective can be defined as:

$$\mathcal{J}_{\text{for}}(\theta_\beta) = \mathbb{E}_{q \sim P(Q), \{m_i\}_{i=1}^{G} \sim \pi_{\theta_\beta^{\text{old}}}(\cdot|q, T^*)} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min\left(\rho_i(\theta_\beta) A_{m_i}, \text{clip}(\rho_i(\theta_\beta), 1-\epsilon, 1+\epsilon) A_{m_i}\right) \Bigg] \tag{9}$$

where $\rho_i(\theta_\beta) = \frac{\pi_{\theta_\beta}(m_i|q)}{\pi_{\theta_\beta^{\text{old}}}(m_i|q)}$ denotes the probability ratio, ϵ is clip hyper-parameters and A_{m_i} is the advantage computed as $\frac{r_m(m_i) - \text{mean}_{k=1}^{G} r_m(m_k)}{\text{std}_{k=1}^{G} r_m(m_k)}$. Similarly, the solving layer maximizes:

$$\mathcal{J}_{\text{sol}}(\theta_\gamma) = \mathbb{E}_{q \sim P(Q), \{o_{ij}\}_{j=1}^{P} \sim \pi_{\theta_\gamma^{\text{old}}}(\cdot|m_i)} \Bigg[\frac{1}{P} \sum_{j=1}^{P} \min\left(\rho_i(\theta_\gamma) A_{o_{ij}}, \text{clip}(\rho_i(\theta_\gamma), 1-\epsilon, 1+\epsilon) A_{o_{ij}}\right) \Bigg] \tag{10}$$

where $\rho_i(\theta_\gamma) = \frac{\pi_{\theta_\gamma}(o_{ij}|m_i)}{\pi_{\theta_\gamma^{\text{old}}}(o_{ij}|m_i)}$ denotes the probability ratio, $A_{o_{ij}}$ is the advantage calculated as $\frac{r_o(o_{ij}) - \text{mean}_{k=1}^{P} r_o(o_{ik})}{\text{std}_{k=1}^{P} r_o(o_{ik})}$. Following the spirit of R1 (Guo et al., 2025), for reward $r_m(\cdot)$ and $r_o(\cdot)$, we also employ a rule-based reward that consists of correctness and format. In the reward function $r_m(m_i)$ for π_{θ_β}, the correctness score is defined as the average reward of outputs sampled by the lower-level model given m_i.

Therefore, the tri-level optimization problem can be finally formulated as:

$$\begin{aligned} \max_{\theta_\gamma, \theta_\beta, \theta_\alpha} \quad & \mathcal{J}_{\text{sol}}(\theta_\gamma, \theta_\beta^*, \theta_\alpha^*) \\ \text{s.t.} \quad & \theta_\beta^* = \arg\max_{\theta_\beta} \mathcal{J}_{\text{for}}(\theta_\gamma, \theta_\beta, \theta_\alpha^*) \\ & \quad \text{s.t.} \quad \theta_\alpha^* = \arg\max_{\theta_\alpha} \mathcal{J}_{\text{meta}}(\theta_\gamma, \theta_\beta, \theta_\alpha) \end{aligned} \tag{11}$$

where $\mathcal{J}(\cdot)$ denotes the objective functions.

5.2 TriRL Algorithm

Algorithm 1 TriRL Algorithm

Initialize: Parameters $\theta_\gamma^{(0)}$, $\theta_\beta^{(0)}$, $\theta_\alpha^{(0)}$; initial penalties $\lambda^{(0)} > 0$, $\mu^{(0)} > 0$; growth factors $\rho_\lambda > 1$, $\rho_\mu > 1$; learning rates η_γ, η_β, η_α; iterations T; update interval K.
for $t = 1$ to T **do**
  Compute $\nabla_{\theta_\alpha} J_{\text{meta}}(\cdot)$.
  Compute $\nabla_{\theta_\beta} J_{\text{for}}(\cdot)$.
  Compute $P_1^{(t-1)}$ and $P_2^{(t-1)}$ using (14) and (15).
  Compute gradient of (16) w.r.t. θ_γ, θ_β, θ_α.
  Update: $\theta_\gamma^{(t)}$, $\theta_\beta^{(t)}$, $\theta_\alpha^{(t)}$.
  if $t \mod K = 0$ **then**
    $\lambda^{(t)} \leftarrow \rho_\lambda \lambda^{(t-1)}, \quad \mu^{(t)} \leftarrow \rho_\mu \mu^{(t-1)}$.
  else
    $\lambda^{(t)} \leftarrow \lambda^{(t-1)}, \quad \mu^{(t)} \leftarrow \mu^{(t-1)}$.
  end if
end for
Return: $\theta_\gamma^{(T)}$, $\theta_\beta^{(T)}$, $\theta_\alpha^{(T)}$.

To address the tri-level optimization problem formulated in (11), we propose the TriRL algorithm, which leverages a penalty function approach to handle the nested structure of the optimization.

The key insight is to relax the argmin constraints by penalizing deviations from the first-order optimality conditions. For the innermost problem, the stationarity condition is:

$$\nabla_{\theta_\alpha} \left(J_{\text{meta}}(\theta_\gamma, \theta_\beta, \theta_\alpha) \right) = 0. \tag{12}$$

Similarly, for the middle-level problem:

$$\nabla_{\theta_\beta} \left(J_{\text{for}}(\theta_\gamma, \theta_\beta, \theta_\alpha) \right) = 0. \tag{13}$$

We define quadratic penalty terms to measure these violations:

$$P_1(\theta_\gamma, \theta_\beta, \theta_\alpha) = \left\| \nabla_{\theta_\alpha} \left(J_{\text{meta}}(\theta_\gamma, \theta_\beta, \theta_\alpha) \right) \right\|^2, \tag{14}$$

$$P_2(\theta_\gamma, \theta_\beta, \theta_\alpha) = \left\| \nabla_{\theta_\beta} \left(J_{\text{for}}(\theta_\gamma, \theta_\beta, \theta_\alpha) \right) \right\|^2. \tag{15}$$

The penalized objective function then becomes:

$$\begin{aligned} \max_{\theta_\gamma, \theta_\beta, \theta_\alpha} \quad & F = J_{\text{sol}}(\theta_\gamma, \theta_\beta, \theta_\alpha) - \\ & \lambda P_2(\theta_\gamma, \theta_\beta, \theta_\alpha) - \mu P_1(\theta_\gamma, \theta_\beta, \theta_\alpha) \end{aligned} \tag{16}$$

where $\lambda > 0$ and $\mu > 0$ are penalty parameters that are increased over iterations to enforce the constraints asymptotically.

The TriRL algorithm (Algorithm 1 proceeds by iteratively optimizing the penalized objective using policy gradient. To ensure convergence, the penalty parameters are adaptively increased, promoting adherence to the inner optimality conditions while optimizing the outer objective.

6 Experiments

Baselines We consider the following baselines, which respectively represent different paradigms: **(a) Standard** CoT (Wei et al., 2022), POT (Chen et al., 2022); **(b) Structured Reasoning** Plan-and-Solve (Wang et al., 2023), InfoRe (Cheng et al., 2024), BoT (Yang et al., 2024b), ToT (Yao et al., 2023); **(c) Vanilla Training** RL_text (fine-tuned to generate CoT-style textual reasoning) and RL_code (fine-tuned to generate Python code), both optimized with the GRPO algorithm.

Evaluation Datasets We conduct extensive experiments on nine datasets spanning diverse TCO problem types, where answers cannot be directly derived from prior knowledge or common sense, including eight established benchmarks and one new synthetic benchmark: GSM_hard, a mathematical benchmark with uncommon large numbers; Cladder and CounterBench for multi-step causal reasoning; K&K puzzles and AutoLogi for different logical puzzles independent of commonsense knowledge; and Next-Step Prediction and ToT for spatial and temporal commonsense reasoning respectively. And to systematically evaluate multi-constraint integration, we additionally construct

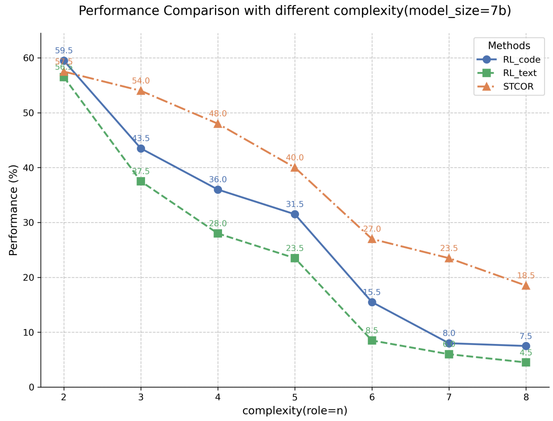


Figure 3: Case study on the K&K dataset: the x-axis represents problem difficulty (2–8 characters) with the relative advantage of STCOR over the suboptimal baselines RL_code and RL_text becoming more pronounced as complexity increases.

TCO-Bench, a synthetic dataset extended from GSM8K (Cobbe et al., 2021), which contains two difficulty subsets: the easy subset mixes two different constraint types, while the hard subset integrates three or more distinct constraint types for comprehensive reasoning.

Base Models and Experimental Setup The framework employs a BERT (Devlin et al., 2019) model for the Meta-Modeling layer, while the Formalization and Solving layers use two model scales **Qwen2.5-1.5B/7B-Instruct** (Yang et al., 2024a). In the training phase, we trained both the vanilla baselines and our algorithm on the GSM8K under the same LoRA rate. For each dataset, we provide a manually crafted template to construct knowledge base. During inference, the meta-modeling layer selects the top $K = 4$ templates for downstream processing. All methods are evaluated under consistent few-shot settings (1-3 shots) with fixed hyperparameters (temperature=0.7, top_p=0.95). No training is performed on evaluation benchmarks except where noted.

6.1 Main results

As shown in Table 1, our STCOR framework demonstrates superior performance across diverse reasoning benchmarks, substantially outperforming all baseline methods. The framework exhibits three key advantages:

Cross-Domain Generalization In contrast to vanilla training approaches, which typically yield superior performance for small models on datasets

Model	Method	Mathematical		Causal		Logical		Commonsense		TCO		
		GSM8K	GSM_hard	Cladder	CB	Autologi	K&K	ToT	Next-Step	easy	hard	Average
qwen2.5-7b-Instruct	COT	88.7	41.5	64.5	63.2	52.5	15.8	59.5	36.5	34.0	34.0	49.0
	POT	88.5	57.5	53.5	67.2	53.0	22.7	67.5	38.8	26.5	23.0	49.8
	PS	85.8	36.5	62.8	57.2	51.0	16.0	59.5	35.0	31.0	35.0	47.0
	InfoRe	88.3	46.5	63.0	65.5	54.5	17.0	62.3	37.8	35.0	29.0	50.0
	TOT	88.9	47.9	63.5	64.6	53.2	17.5	62.6	36.9	34.0	35.0	50.4
	BOT	89.1	54.3	64.0	66.1	53.8	19.3	65.7	37.4	31.5	29.0	51.0
	RL_text	89.3	49.0	62.2	70.0	55.8	22.3	61.0	37.0	37.5	37.0	52.1
	RL_code	89.0	62.5	56.0	72.2	54.2	27.5	70.0	40.5	32.0	23.0	52.7
	STCOR	**90.5**	**64.3**	**69.0**	**82.5**	**62.0**	**36.8**	**78.5**	**48.3**	**47.5**	**41.0**	**62.0**
qwen2.5-1.5b-Instruct	COT	54.0	17.5	53.8	50.2	26.5	4.5	15.0	17.5	13.0	11.0	26.3
	POT	48.5	38.5	43.8	47.3	24.5	4.3	21.0	27.8	8.0	7.0	27.1
	PS	59.3	17.5	49.0	47.5	26.0	4.8	12.5	16.0	11.0	11.0	25.5
	InfoRe	52.8	17.8	54.5	52.0	24.8	5.0	17.8	15.5	11.0	8.0	26.0
	TOT	55.8	23.8	54.0	50.8	26.2	4.7	17.2	20.7	12.0	10.0	27.5
	BOT	57.5	30.2	54.3	51.4	26.3	4.8	19.3	23.8	11.0	9.0	28.8
	RL_text	58.5	21.0	52.5	51.25	26.7	5.3	20.3	21.5	13.0	12.0	28.2
	RL_code	60.5	44.0	47.25	54.0	23.3	5.8	30.0	30.0	9.5	11.0	31.5
	STCOR	**61.3**	**45.0**	**58.0**	**63.0**	**29.4**	**9.5**	**37.5**	**35.3**	**19.5**	**13.0**	**37.2**

Table 1: We report the average accuracy(%) for 2 runs for each dataset on four types of reasoning tasks. Among them, GSM8K is an in-domain dataset, and the rest are out-of-domain datasets. **Bold** indicates the best, underline the second-best performance.

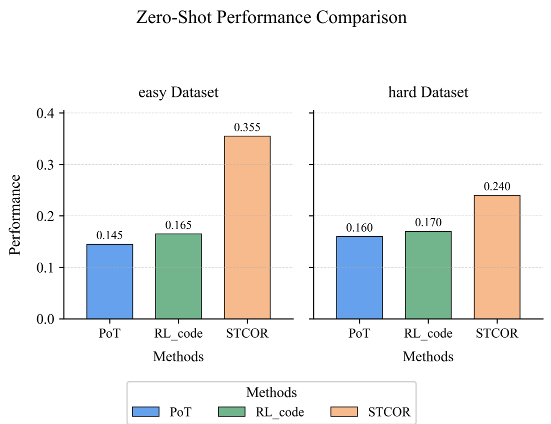


Figure 4: Performance on TCO-Bench under zero-shot settings with qwen2.5-7B model.

Better Performance on Complex Problems Our method also exhibits marked improvements on complex problems. Compared to suboptimal baselines, the 7B STCOR achieves a maximum 33% performance gain on the K&K dataset, which has the lowest standard CoT accuracy, indicating higher difficulty. Similarly, the 1.5B STCOR attains a 60% maximum advantage on K&K and a 16.7% improvement on CB. Notably, CB is a binary classification dataset where 50% CoT accuracy equals random selection, signifying substantial complexity. Crucially, on our multi-constraint integration-focused TCO-Bench, the 7B STCOR outperforms the best baseline by 26.7% on the Easy subtask. A case study in Figure 3 (K&K dataset) further shows that our method's performance advantage grows with increasing problem complexity.

resembling the training set (e.g., GSM_hard) or those requiring analogous sequential reasoning (e.g., CounterBench), STCOR demonstrates enhanced cross-domain generalization. It achieves average performance gains of 17.6% and 18.1% over the second-best baseline (RL_code) for the 7B and 1.5B models respectively. This cross-domain robustness stems from STCOR's hybrid architecture, where optimization-form models capture universal reasoning patterns while formalization uncovers implicit structural knowledge.

Zero-Shot Performance Analysis As shown in the figure 4, we conduct additional zero-shot experiments on the TCO datasets. Since the composite constraint patterns in TCO problems are relatively uncommon in pre-training data, directly prompting the base model (e.g., zero-shot PoT) yields limited performance. In contrast, STCOR achieves markedly higher zero-shot accuracy due to its syllogistic reasoning structure and guidance from prototype templates. This demonstrates that STCOR's structured reasoning pattern and knowl-

(a) Qwen2.5-7B-Instruct

Variant	CB	K&K	TCO	Mean Δ
STCOR (full)	**82.5**	**36.8**	**47.5**	–
w/o Meta	80.3	32.5	45.5	−2.83
w/o Formal	76.8	31.2	42.0	−5.60
w/o Solving	65.2	18.7	38.5	−14.80
w/o Train	75.8	30.5	43.0	−5.80

(b) Qwen2.5-1.5B-Instruct

Variant	CB	K&K	TCO	Mean Δ
STCOR (full)	**63.0**	**9.5**	**19.5**	–
w/o Meta	58.0	7.0	18.5	−2.83
w/o Formal	53.5	5.0	12.0	−7.20
w/o Solving	54.0	5.5	12.5	−6.60
w/o Train	55.5	6.0	14.5	−5.30

Table 2: Performance of STCOR and its ablated variants on three representative datasets. "Mean Δ" denotes the average performance drop across the three datasets relative to the full STCOR.

edge base can effectively unleash the potential of code-augmented reasoning.

6.2 Ablation Study

We conducted ablation experiments using three typical datasets: CounterBench, K&K, and our TCO-Bench. Five variants were considered: (1) **STCOR**: our full framework; (2) **w/o Meta**: removing meta-modeling layer; (3) **w/o Formal**: removing formalization layer; (4) **w/o Solving**: solving with standard CoT. (5)**w/o train**: the full framework without TriRL training. Results show that removing any component causes performance degradation (2-18 points across datasets), with solving ablation causing the largest drop. The results are presented in Table 2.

7 Limitations

While STCOR provides a principled reasoning framework, it currently relies on manually curated templates, which requires significant expert effort and limits scalability to new domains. Future work will explore dynamic template generation through self-iterative learning and few-shot template adaptation to reduce this engineering burden and enhance framework generalization.

8 Ethical Considerations

The data for the proposed methods is drawn solely from publicly accessible project resources on reputable websites, ensuring that no sensitive information is included. Scaling compute may result in substantial electricity consumption and carbon dioxide emissions. In practical scenario applications, data privacy issues must be taken into account.

9 Conclusion

In this work, we present STCOR, a syllogism-driven reasoning framework that transforms textual constrained optimization (TCO) from step-wise exploration into principled, first-principles instantiation. By decomposing reasoning into three structured levels—meta-modeling, formalization, and solving—and supporting them with the TriRL joint training algorithm, this work makes several key contributions: (i) it introduces a novel TCO problem formulation and corresponding STCOR reasoning paradigm, departing from prior step-based methods by emphasizing principled formalization to capture problem essence; (ii) the tri-level architecture ensures full transparency and auditability, with each reasoning step grounded in explicit logical premises; (iii) extensive experiments demonstrate significant performance gains while providing naturally interpretable reasoning paths. STCOR represents a promising step toward building more trustworthy, systematic, and principled reasoning systems with large language models.

10 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 12371519 and 61771013; in part by the Shenzhen Loop Area Institute under project FP202602.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, and Piotr Nyczyk. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Xiaoxia Cheng, Zeqi Tan, Wei Xue, and Weiming Lu. 2024. Information re-organization improves reasoning in large language models. *Advances in Neural Information Processing Systems*, 37:130214–130236.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, and Reiichiro Nakano. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. *In The Eleventh International Conference on Learning Representations*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and Xiao Bi. 2025. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Ruixin Hong, Hongming Zhang, Xiaoman Pan, Dong Yu, and Changshui Zhang. 2024. Abstraction-of-thought makes language models better reasoners. *arXiv preprint arXiv:2406.12442*.

Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*.

Yancheng Huang, Kai Yang, Zelin Zhu, and Leian Chen. 2024. Triadic-OCD: Asynchronous Online Change Detection with Provable Robustness, Optimality, and Convergence. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 238 of *Proceedings of Machine Learning Research*, pages 20382–20412. PMLR. Article No. 819.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, and Alex Carney. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Chengtao Jian, Kai Yang, and Yang Jiao. 2024. Tri-level navigator: LLM-empowered tri-level learning for time series ood generalization. *Advances in Neural Information Processing Systems*, 37:110613–110642.

Chengtao Jian, Kai Yang, Ye Ouyang, and Xiaozhou Ye. 2025. Stable preference optimization for LLMs: A bilevel approach beyond direct preference optimization. *arXiv e-prints*, pages arXiv–2507.

Yang Jiao, Kai Yang, and Chengtao Jian. 2025. Dtzo: Distributed trilevel zeroth order learning with provable non-asymptotic convergence. In *Forty-second International Conference on Machine Learning*.

Yang Jiao, Kai Yang, Tiancheng Wu, Chengtao Jian, and Jianwei Huang. 2024. Provably convergent federated trilevel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12928–12937.

Junlong Li, Daya Guo, Dejian Yang, Runxin Xu, Yu Wu, and Junxian He. 2025. Codei/o: Condensing reasoning patterns via code input-output prediction. *arXiv preprint arXiv:2502.07316*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Sohini Roychowdhury. 2024. Journey of hallucination-minimized generative ai solutions for financial decision makers. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1180–1181.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, and Y Wu. 2024. Deepseekmath:

Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Kaya Stechly, Matthew Marquez, and Subbarao Kambhampati. 2023. Gpt-4 doesn't know it's wrong: An analysis of iterative prompting for reasoning problems. *arXiv preprint arXiv:2310.12397*.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. 2024. Toward self-improvement of LLMs via imagination, searching, and criticizing. *Advances in Neural Information Processing Systems*, 37:52723–52748.

Ziyu Wan, Yunxiang Li, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, and Ying Wen. 2025. Rema: Learning to meta-think for LLMs with multi-agent reinforcement learning. *arXiv preprint arXiv:2503.09501*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Zhiyuan Wang, Bokui Chen, Yinya Huang, Qingxing Cao, Ming He, Jianping Fan, and Xiaodan Liang. 2025. Ormind: A cognitive-inspired end-to-end reasoning framework for operations research. *arXiv preprint arXiv:2506.01326*.

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.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, and Dakota Mahan. 2025. Towards system 2 reasoning in LLMs: Learning how to think with meta chain-of-though. *arXiv preprint arXiv:2501.04682*.

Ziyang Xiao, Dongxiang Zhang, Yangjun Wu, Lilin Xu, Yuan Jessica Wang, Xiongwei Han, Xiaojin Fu, Tao Zhong, Jia Zeng, and Mingli Song. 2023. Chain-of-experts: When LLMs meet complex operations research problems. In *The twelfth international conference on learning representations*.

Wantong Xie, Yi-Xiang Hu, Jieyang Xu, Feng Wu, and Xiangyang Li. Murka: Multi-reward reinforcement learning with knowledge alignment for optimization tasks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN international symposium on machine programming*, pages 1–10.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Haoran Wei. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. 2024b. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems*, 37:113519–113544.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. 2023. Satlm: Satisfiability-aided language models using declarative prompting. *Advances in Neural Information Processing Systems*, 36:45548–45580.

Junchi Yu, Ran He, and Rex Ying. 2024. Thought propagation: An analogical approach to complex reasoning with large language models. *In The Eleventh International Conference on Learning Representations*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, and Quoc Le. 2024a. Least-to-most prompting enables complex reasoning in large language models. *In The Eleventh International Conference on Learning Representations*.

Jin Peng Zhou, Charles Staats, Wenda Li, Christian Szegedy, Kilian Q Weinberger, and Yuhuai Wu. 2024b. Don't trust: Verify–grounding LLM quantitative reasoning with autoformalization. *In The Eleventh International Conference on Learning Representations*.

Yu Zhu and Kai Yang. 2026. Llm-driven multi-turn task-oriented dialogue synthesis for realistic reasoning. *arXiv preprint arXiv:2602.23610*.

Beyond Logical Forms: LLM-Extracted Patterns for Fallacy Classification

Eleni Papadopulos[1,2], **Firoj Alam**[3] **Giovanni Da San Martino**[2]
[1]Politecnico di Torino, Italy, [2]Università di Padova, Italy
[3]Qatar Computing Research Institute, Qatar
eleni.papadopulos@polito.it, giovanni.dasanmartino@unipd.it
fialam@hbku.edu.qa

Abstract

In today's fast-paced information era, logical fallacies, defined as defective patterns of reasoning, inevitably contribute to the growth of information disorder. However, often fallacies appear in nuanced forms that complicate automated classification. In this study, we investigate whether merging abstract logical structures with context-level linguistic cues proves beneficial for fallacy classification, developing a framework that inductively extracts such patterns from fallacious examples and their explanations using Large Language Models (LLMs). We evaluate the impact of these patterns across different LLMs and experimental zero- and one-shot configurations, showing statistically significant improvements over zero-shot baselines and outperforming competing approaches. Cross-dataset experiments validate generalization, establishing data-driven pattern extraction as an effective method for generating logical representations.

1 Introduction

A logical fallacy is a common thinking error, especially one apt to mislead (Gensler, 2010). These arguments often appear rational and logically coherent on the surface, but deeper analysis reveals they are not (Copi et al., 1953). Fallacies are traditionally classified into formal and informal types: ***formal fallacies*** violate the rules of logical structure regardless of content, while ***informal fallacies*** are patterns of mistakes that are made in the everyday uses of language and are related to contextual meaning (Hamblin, 1970; Bacon et al., 1999).

To evaluate the quality of an argument, it is helpful to reconstruct it into what is known as logical form, the structure that emerges when the specific content of a statement is replaced by variables (Johnson and Blair, 1977). For example, the argument *If it rains, then the ground will be wet. It is raining. Therefore, the ground is wet* has the logical form *If P, then Q. P. Therefore, Q*. Building on this formalization framework, Jin et al. (2022) developed a structure-aware model for fallacy detection on the LOGIC dataset that compares arguments' and fallacies' logical forms. However, in their approach a single logical form is assigned to each fallacy, which might fail to capture the full spectrum of ways a fallacy can manifest in natural discourse. Another challenge is related to informal fallacies, where reasoning is often more nuanced and context-dependent than abstract representations suggest.

These limitations motivate the need to go beyond purely abstract representations, incorporating linguistic elements, such as lexical markers or rhetorical devices, to provide a more comprehensive characterization of how fallacies manifest in natural language. We argue that LLMs can inductively extract such representations from fallacious examples, capturing both the logical structure and linguistic cues that reveal the underlying mechanisms of deception. Unlike prior work that formalized fallacy logic through hand-crafted templates (Robbani et al., 2024), our approach is data-driven and not restricted to a limited number of fallacies. Hereafter, we refer to these extracted structures collectively as *structural patterns*. Our goal is to investigate whether context-aware structural information is valuable for automated fallacy detection.

While existing supervised approaches require heavy computational resources for fine-tuning (Vijayaraghavan and Vosoughi, 2022; Lei and Huang, 2024; Sourati et al., 2023a,b; Alhindi et al., 2024), to our knowledge, no prior work has explored fallacy classification from a structural perspective without any additional fine-tuning. Although we use labeled data for pattern extraction (resulting in a weakly supervised approach), our framework avoids fine-tuning costs and produces generalizable patterns that allow classification through prompting alone, enabling comparison with unsu-

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 12–25

pervised methods.

We evaluate multiple prompting configurations to determine which components enhance performance and examine the impact of demonstrations on detection capabilities. Our approach, incorporating generated patterns, achieves noteworthy results among unsupervised methods on the dataset LOGIC. Finally, to validate the robustness and transferability of our patterns, we assess their performance across two further datasets spanning diverse domains and argumentative styles.

In summary, our contributions are threefold:

- We leverage LLMs to automatically extract patterns from fallacious examples and their explanations, which are then employed in inference-only classification.
- We evaluate different LLMs with various prompt designs outperforming competing approaches on the LOGIC dataset.
- We validate generalizability by testing our patterns on two different datasets with different domains and structures.

2 Related Work

Recent advances in fallacy detection have increasingly turned to LLMs, though few studies have relied exclusively on prompting-based techniques. Several works have employed fallacy detection to probe LLMs' logical reasoning abilities (Teo et al., 2025; Hong et al., 2024; Li et al., 2024; Xu et al., 2026). Among these, Hong et al. (2024) investigated self-verification capabilities and showed that LLMs face more challenges with structure-based (formal) fallacies with respect to content-based (informal) ones, and that fallacy definitions provide minimal improvements. Xu et al. (2026) has shown that reasoning models have better performances with respect to non-reasoning ones for fallacy classification. Among studies relying exclusively on prompting techniques, Pan et al. (2024) designed single-round and multi-round prompting schemes for zero-shot detection, while Jeong et al. (2025) introduced contextual prompting incorporating counterarguments, explanations, and goals with confidence-based ranking, showing that explanations particularly enhance performance. Lim and Perrault (2024) assessed detection abilities on the LOGIC dataset using few-shot prompting, though their different taxonomy limits direct comparison with our work. Other research has examined the logical structure of argumentation. Most notably, Jin et al. (2022) developed a structure-aware model based on Electra that distills arguments into logical forms and compares them against fallacy patterns sourced from logicallyfallacious.com. Another prominent framework in this field is Walton's theory of argumentation (Walton, 2008), consisting in about 60 templates that capture common argument types, each associated with a set of critical questions to evaluate their validity. These schemes have been adopted in computational approaches for disinformation (Gutiérrez-Mandingorra et al., 2024), misinformation (Ruiz-Dolz and Lawrence, 2025) and fake news (Wang et al., 2025) detection. Regarding fallacies specifically, Ruiz-Dolz and Lawrence (2023) introduced a dataset of sentences grounded in Walton's argumentation schemes, labeling them as fallacies when the associated critical questions could not be successfully answered. However, Walton's schemes do not hold a one-to-one correspondence with fallacies and cover only a limited number of them, limiting their direct applicability to fallacy classification. Of particular relevance is the work of Robbani et al. (2024), who re-designed four of Walton (2008) and Reisert et al. (2018)'s schemes with the goal of expliciting fallacies' implicit logic, introducing formal logical schemas with explicit variables and relationships. While this represents a meaningful step toward structural formalization, their patterns are manually designed and leave a large portion of fallacy types unrepresented. Our patterns, by contrast, are extracted automatically from data, allowing to adapt to intra-class variation.

3 Datasets

The LOGIC dataset is a collection of 2,449 examples across 13 fallacy types (Jin et al., 2022). Instances are sourced from educational platforms about fallacies such as Quizziz and study.com. The dataset consists of brief dialogues and short statements. Given the educational intent behind these examples, sentences tend to have relatively straightforward syntactic structures, making the dataset particularly well-suited for the extraction of the patterns.

Although it contains 13 distinct classes, a thorough analysis revealed that some of the classes actually contain instances of different fallacies, that were grouped together. For instance, the class *Hasty Generalization* contains examples of actual

Class	*Fallacies included*
Intentional Fallacy	Intentional Fallacy Shifting the Burden of Proof Moving the Goalposts No True Scotsman
False Cause	Post Hoc False Cause
Hasty Generalization	Hasty Generalization Slippery Slope

Table 1: Examples of classes in LOGIC dataset containing instances of different fallacy types. While coherent, these groupings comprise fallacies with distinct structural patterns. A detailed breakdown of all classes' subtypes is provided in Appendix D of Jin et al. (2022).

Hasty Generalization as well as *Slippery Slope* (Table 1). While these grouped fallacies share common logical flaws and thus belong to the same conceptual group, they manifest through different structural patterns.

We experiment on two further datasets: REDDIT (Sahai et al., 2021), consisting of fallacious comments extracted from subreddits covering different topics and ELECDEBATE60TO16 (hereafter ELECDEBATE) (Goffredo et al., 2023), a collection of televised debates of the presidential election campaigns in the U.S. from 1960 to 2016. Some fallacy classes contain sub-categories. In Table 2, we report a summary of the dataset, and a description of each taxonomy is provided in Appendix B.

Data	Dataset split	# Classes	Genre	Domain
LOGIC	1807/299/299	13	Dialogue	Education
REDDIT	588/148/105	8‡	Comments	General
ELECDEBATE	1120/200/187	6	Dialogue	Politics

Table 2: Statistics of the three datasets. ‡ indicates that the *No Fallacy* class is included.

4 Pattern Generation

Natural arguments appear in several forms. Such variability manifests itself in LOGIC dataset as well as many others (Habernal et al., 2018; Da San Martino et al., 2019). For this reason, we address our research question by modeling patterns inductively from the training set of LOGIC. The choice of the dataset for pattern extraction is critical. It provides the required combination of structural clarity and fallacy diversity through its multiple sub-types per class. These properties make it especially suited for our purpose. The clean argumentative structure allows to formalize clear logical patterns while capturing intra-class variations.

Our pattern generation procedure features two steps:

Step 1: Explanation Generation Explanations have been shown to be instrumental in identifying and discrediting fallacious reasoning, as they make the logical structure of arguments explicit and open to scrutiny (Storer, 1949). Furthermore, Jeong et al. (2025) has demonstrated that providing explanations constitutes valuable contextual information in zero-shot settings. We expected explanations to facilitate pattern extraction by breaking down the reasoning process and revealing shared reasoning flaws, particularly useful for informal fallacies.

Given a sentence from the training set and its fallacy label, we used `llama-3.3-70B-Instruct` (Dubey et al., 2024) to generate an explanation that justifies why that sentence contains the specified fallacy.

Step 2: Pattern Extraction For each fallacy class, we used OpenAI's reasoning model `o4-mini` (OpenAI, 2025) to extract patterns from all the sentences of that class and their explanations, requiring the model to preserve function words such as prepositions or adverbs and to abstract away from content words by replacing them with placeholders while keeping the original reasoning form. Additionally, summaries were extracted to derive new fallacy definitions.

We opted for `llama-3.3-70B-Instruct` for explanation generation as it provided high-quality explanations while remaining cost-effective for large-scale text generation. For pattern extraction, we employed `o4-mini` given its reasoning capabilities. The prompts used in our experiments are reported in the Github repository.[1]

In the initial phase of our research, we aimed to cover two distinct logical aspects from our arguments and explanations, specific to formal and informal fallacies, respectively:

- arguments' **logical structure** inspired by formal logic theory;
- recurring **lexical schemes** that frequently appear in both sentences and explanations, capturing specific information about the reasoning behind the fallacy as well as frequent syn-

[1] https://github.com/elenipapadopulos/fallacy-patterns

tactic particles, phrases, and examples that convey the fallacious intent.

Our patterns incorporate both of these aspects, as Table 3 shows. These LOGIC-based patterns combine reasoning structure (variables X, Y, Z) with concrete linguistic features (specific phrases, loaded terms, rhetorical devices), occasionally retaining some definitions. The full list of patterns is available in the repository.[1]

The process resulted in approximately 3-6 patterns per fallacy class. Final patterns were obtained after providing different subsets to the model and selecting the best performing one on the validation set, in the attempt to retain only useful information and avoid redundancy. In section 3 we discussed how one class in the datasets could correspond to multiple fallacies. Although in some cases, e.g. a pattern for *Tu quoque* (a fallacy which is part of the class *Ad Hominem* in LOGIC), is correctly generated and selected, sometimes fails to select patterns when multiple fallacies are grouped under the same class label. This is expected because we include the fallacy class name in the prompt, which likely biases the model toward patterns that match its internal knowledge of that particular class name. To ensure a broader coverage of fallacies listed in Table 1, we manually isolated instances of frequent and undetected fallacies (such as *Shifting the Burden of Proof*) and repeated the procedure.

Intentional Fallacy Patterns
1. The argument assumes that because X (e.g., someone's intention, belief, or lack of counter-evidence), therefore Y is true.
2. Asserting P is true because it has not been disproven.
3. Because the creator intended [interpretation], the work should be understood as [interpretation].
4. Questions framed to presuppose guilt or a specific intention (e.g., "Have you stopped X?"), thus assuming what is to be proven.
5. If A does not have trait X, and X is allegedly typical of group G, then A is not a member of G.
Red Herring Patterns
1. Instead of addressing [original issue], the argument shifts focus to [irrelevant topic], which distracts from the main discussion.
2. The argument attempts to justify, explain, or defend by referencing [irrelevant detail], ignoring the original issue of [main topic].
3. A shift from the initial question or problem to a secondary topic that does not logically follow, e.g., "You asked about X, but I will tell you about Y."

Table 3: Patterns for *Intentional Fallacy* and *Red Herring*. For *Intentional Fallacy*, patterns (#4) and (#5) illustrate lexical schemes and logical forms, respectively, that encode intent and structure.

5 Experiments

This section describes our experiments for fallacy classification, including our patterns extracted by the procedure introduced in Section 4 and several competing prompting strategies. Additional experiments are reported in Appendix D. We used the following LLMs for our experiments: `gpt-4o`, `o4-mini`, `gpt-4.1-mini`, `LLama-3.3-70B`, `deepseek-r1` and `Gemma-3-27B-it` for a total cost of 75 USD. Our intent was to test LLMs from different providers and with different sizes and to compare reasoning and non-reasoning models.

5.1 Prompt Design

Baselines. We compared our approach against several baselines that vary in the type and amount of information provided to the model. The simplest baseline (**ZERO-SHOT**) provides only the list of fallacy names in the dataset as a reference, establishing a minimal information condition. Our second baseline incorporates fallacy definitions to provide more comprehensive background knowledge (**DEF**). These definitions were initially sourced from Lei and Huang (2024) and subsequently refined based on our analysis to ensure clarity and consistency. Finally, we tested a baseline using standard logical forms, following the approach of Jin et al. (2022) and sourcing these forms from logicallyfallacious.com. This final baseline (**LOGICAL FORMS**) allows us to assess the effectiveness of expert-made logical representations compared to our generated pattern-based approach.

LLM-derived Patterns and Definitions. Beyond generating structural patterns, we leveraged the explanations from Section 4 to automatically create new fallacy definitions based on LOGIC training samples. We then replicated experiment **DEF** with these new definitions (**NEW DEF**). We also exploited the patterns extracted by adding them to the prompt (**PATTERNS**) and by implementing a two-step approach where we first ask the LLM to identify the pattern and then to output the corresponding fallacy (**PATTERN MATCHING**).

One-shot Prompting. We further investigated the impact of providing examples to the model through several experimental configurations (Brown et al., 2020), with one-shot prompting proving most effective. Initially, we tested a static approach where one example per fallacy was randomly selected and shown to all test sentences (**ONE-SHOT**), establishing a baseline for example-based learning. To enhance this approach, we augmented the same examples with manually crafted explanations following our previously established definitions as guidelines (**ONE-SHOT + EXP**). We sampled 5 different example sets and performance across all configurations was assessed over 5 runs to ensure statistical reliability.

More sophisticated was our dynamic one-shot prompting approach (**DYNAMIC ONE-SHOT**), which computes embeddings for both training and test sentences to retrieve, for each test sentence, the most similar example per class in the training set. We used `sentence-transformers/all-MiniLM-L6-v2`[2] model and `cross-encoder/stsb-roberta-base`[3] cross-encoder from SentenceTransformers (Reimers and Gurevych, 2019) to compute embeddings and employed cosine similarity to evaluate similarity. We included the previously generated explanations of examples in the prompt as well (**DYNAMIC + EXP**).

Furthermore, we explored structure-focused similarity. Since Jin et al. (2022) released a version of LOGIC with masked arguments (with content words replaced by placeholders), we conducted the same similarity-based procedure using these masked sentences (see an example in Table 4) in the attempt to force the embedding model to focus on structural rather than lexical similarities. For this configuration (**SYNTAX-BASED DYNAMIC ONE-SHOT**), we used `sentence-transformers/all-MiniLM-L6-v2` from SentenceTransformers alongside a syntax-augmented version of RoBERTa-large extracted from Sachan et al. (2021) (see Appendix E).

Finally, we incorporated the generated patterns into our dynamically retrieved examples and their explanations (**DYNAMIC + EXP + PATTERNS**).

Multi-step Classification. An alternative approach decomposes the classification task into three sequential steps within a single model call (**MULTISTEP**) using chain-of-thought prompting (Wei et al., 2022). In the first step, the model is required to generate a structural pattern from the argument according to predefined structural rules. Subsequently, the model should match it to one of the patterns and, as a result, classify the argument.

Original argument	Every time I wear this necklace, I pass my exams. Therefore, wearing this necklace causes me to pass my exams.
Masked argument	Every time MSK<0> MSK<2>, MSK<0> MSK<4>. Therefore, MSK<2> causes MSK<0> to MSK<4>.

Table 4: Example of a masked argument in LOGIC. The distillation algorithm is explained in Jin et al. (2022). The masked version of the dataset was publicly released by the authors and was not created by us.

5.2 Results and discussion

Table 5 summarizes all experimental configurations and results on LOGIC. It reveals a consistent improvement when the model leverages information about the underlying logic extracted through the LLMs, especially with reasoning models and `gpt-4o`. When using reasoning models, the model-generated definitions yield a 4.65% accuracy improvement over our manually corrected definitions. In the same way, including our generated patterns causes a 8.2% increase with respect to the logical forms extracted by the website logicallyfallacious.com and used in Jin et al. (2022). McNemar's test proved statistical significance for all models using **PATTERNS** against **ZERO-SHOT** and for all except `llama` and `deepseek` against **LOGICAL FORMS** method. When it comes to non-reasoning models, the new definitions do not really affect the performance, whereas using our patterns improves the accuracy by 5.8% on average. For comparison, we test our method against Robbani et al. (2024)'s templates: our patterns outperform said schemes by an average 10.7% across all models.

A notable result is the performance increase achieved through dynamic one-shot prompting. In particular, **DYNAMIC ONE-SHOT** approach (using `all-MiniLM-L6-v2`) yields an average 8.87% increase in accuracy compared to **ONE-SHOT**, despite relying on semantic similarity for example selection. On the other hand, the syntax-oriented example retrieval strategy (**SYNTAX-BASED DYNAMIC ONE-SHOT**) does not outperform the semantic selection. This may be partially due to inaccuracies in the sentence masking process,

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://huggingface.co/cross-encoder/stsb-roberta-base

Method	o4-mini		gpt-4o		deepseek-r1		gpt-4.1-mini		llama-3.3-70B		gemma-3-27b-it	
	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
Baselines												
ZERO-SHOT	61.7	55.3	62.7	57.0	62.7	57.3	**57.8**	**51.0**	55.8	47.7	60.5	51.3
DEF	62.1	58.7	65.0	58.7	62.2	56.5	57.5	50.6	59.1	51.5	**63.5**	**55.2**
LOGICAL FORMS	**63.2**	**57.4**	**65.4**	**59.4**	**63.1**	**55.4**	57.8	49.4	**60.2**	**51.3**	62.8	53.9
LLM-derived Patterns and Definitions												
NEW DEF	66.8	67.3	66.8	59.9	66.8	60.0	57.5	52.5	58.8	53.3	64.8	57.7
PATTERNS	**72.2**	**66.4**	73.2	64.9	70.5	66.2	63.5	55.7	64.5	53.3	**68.5**	**61.9**
PATTERN MATCHING	70.1	65.9	**73.5**	**66.5**	**71.5**	**66.5**	**65.2**	**57.9**	**66.2**	**59.6**	67.2	59.9
One-shot prompting												
ONE-SHOT	63.6	59.6	64.1	58.7	58.5	55.9	56.2	48.1	56.1	46.2	60.0	49.7
ONE-SHOT + EXP	65.2	59.5	63.5	59.0	45.7	48.6	56.8	50.0	56.3	47.9	59.2	49.7
DYNAMIC ONE-SHOT												
all-MiniLM-L6-v2	70.2	67.6	71.3	66.4	70.4	66.2	65.8	61.7	65.5	59.7	68.5	63.3
roberta-base	69.5	64.3	69.5	64.6	72.5	67.8	65.5	61.3	64.8	58.9	66.5	60.6
SYNTAX-BASED DYNAMIC ONE-SHOT												
all-MiniLM-L6-v2	68.2	63.6	71.2	66.1	68.5	64.2	63.2	58.3	62.8	55.7	64.5	57.3
syntax-augmented roberta-large	65.5	65.5	71.2	66.3	68.5	64.2	64.5	58.6	64.2	56.5	63.5	56.0
DYNAMIC + EXP	71.2	68.9	69.5	65.0	72.7	67.9	**67.8**	**61.0**	**67.5**	**62.2**	68.2	63.2
DYNAMIC + EXP + PATTERNS	**74.2**	**68.9**	**73.1**	**67.2**	**73.2**	**67.9**	66.8	62.3	67.2	55.1	**70.5**	**65.9**
Multi-step classification												
MULTISTEP	65.4	64.9	70.9	62.7	62.2	57.1	65.8	55.8	62.5	55.1	66.8	60.2

Table 5: Fallacy classification performance on LOGIC. **Bold**: best approach in section per model by accuracy, **Bold**: best approach overall per model by accuracy. F_1 score denotes Macro F_1 score, which accounts for the class imbalance in the dataset.

which can negatively impact the retrieval of similar examples and the classification, consequently. The **MULTISTEP** approach shows weaker performance than **PATTERN MATCHING**, especially for `deepseek-r1` (DeepSeek-AI et al., 2024), implying that generating logical forms without explicit guidance constitutes the main challenge for the model in the request.

In summary, including context-aware logical patterns proves consistently beneficial for fallacy classification: **PATTERNS** with `gpt-4o` reaches 73.5% accuracy, outperforming prior unsupervised methods (Table 6), while **DYNAMIC+EXP+PATTERNS** with `o4-mini` achieves 74.2% when augmented with examples and patterns.

Method	Acc	F_1
Jeong et al. (2025)	49.0	37.0
Pan et al. (2024)	-	50.5
PATTERNS (`gpt-4o`)	73.5	66.5
DYNAMIC+EXP+PAT. (`o4-mini`)	74.2	68.9

Table 6: Comparison of our best results against the unsupervised baselines provided by Jeong et al. (2025) and Pan et al. (2024) (described in Appendix C) for LOGIC.

5.3 Error analysis

Pattern matching Requesting the model to identify the closest pattern for each argument provides insight into the association process between sentences and patterns. For our analysis, we have split our fallacies into two groups in Table 7: *(i)* group 1, consisting of fallacies whose patterns include logical forms while still including additional contextual cues; *(ii)* group 2, consisting of fallacies that lack highly structured patterns and rely more on contextual and semantic features of the sentence.

Figure 1 shows consistently superior accuracy

Group 1	Group 2
• Ad Hominem • Ad Populum • Circular Reasoning • Irrelevant Authority • False Cause • Hasty Generalization • Deductive Fallacy • Black-and-White Fallacy	• Red Herring • Equivocation • Emotional Language • Extension Fallacy • Intentional Fallacy

Table 7: Grouped fallacy classes based on pattern features for analytical purposes.

for Group 1, whose classes maintain relatively high performance across all experimental settings. The class *Circular Reasoning* emerges as the most accurately predicted class across all models. For what concerns Group 2, the overall accuracy is, on average, 22% lower with respect to Group 1. The classes *Emotional Language*, *Red Herring* and *Extension Fallacy* achieve moderate prediction accuracy, whereas only *Evading the Burden of Proof*'s patterns within the *Intentional Fallacy* category are correctly classified, and *Equivocation* remains entirely undetected by `gpt-4.1-mini` (OpenAI, 2023). In summary, the models achieve better performance on logical fallacies that exhibit clearer structural characteristics but face difficulties with fallacies requiring more nuanced semantic understanding and contextual analysis.

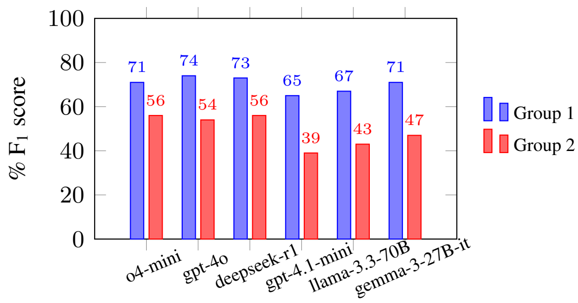


Figure 1: Group-wise F_1 score for each model, relative to the **PATTERN MATCHING** prompt setting.

Furthermore, matching patterns allows us to see that some instances can be deemed as fitting from a structural point of view, thus partially explaining the inherent difficulty of the classification task. While providing guidance through syntactic and logical structure proves beneficial for fallacy detection, this approach does not eliminate all sources of ambiguity, as some sentences may conform to multiple structural patterns. The critical point lies in context-aware pattern application. Models must not only identify logical forms but also evaluate their contextual validity in each sentence.

To quantify the degree of ambiguity inherent in pattern matching, we instructed the best-performing model `o4-mini` to return the five most similar patterns for each argument. This multi-candidate approach enables us to analyze whether lower-ranked patterns might also represent valid interpretations of the same argument. By examining the distribution of pattern similarities and evaluating classification accuracy when considering alternative matches, we can better understand the boundaries of pattern-based classification and identify instances where structural ambiguity genuinely complicates fallacy detection.

Table 9 shows that, when the model is prompted to return multiple matching patterns rather than a single best match, its confidence in the initial prediction decreases, resulting in a 3.4% drop in accuracy (see Table 5).

Acc@1	Acc@2	Acc@3	Acc@4	Acc@5
66.7	75.1	81.8	86.5	88.5

Table 9: Performance analysis in **PATTERN MATCHING** with expanded solution pool: classification results including top 5 predictions as correct.

However, this apparent degradation is misleading when viewed in isolation. By incorporating the second-ranked pattern choice into our evaluation, performance recovers to 75.1%, and continues to improve as we expand our candidate pool to include progressively lower-ranked options. Table 8 illustrates a representative case where the model successfully identifies the correct pattern as its second choice, while its first-ranked selection remains structurally plausible. The model likely assigns one of the *Ad Populum* patterns because it closely matches the argument's logic, while the *Irrelevant Authority* pattern does not fit the sentence since it requires discussion of an unrelated topic, which is not present in the sentence. These subtle distinctions likely make pattern matching more challenging than direct classification because it requires strict structural alignment as well as capturing broader content-related features.

Multistep classification. The **MULTISTEP** approach fails to produce significant results. We conduct this experiment in a single passage to force the model to reason using both semantic and syntactic information. However, classification performance depends critically on the quality of the extracted logical forms, which proves inconsistent and model-dependent. For instance, `o4-mini` embeds classification-relevant contextual information directly into its generated logical forms (Table 10). Furthermore, models demonstrate substantially weaker performance on Group 2 sentences compared to Group 1, showing an average decrease of 21.5% in F_1 score. Additionally, models frequently bypass the pattern matching phase entirely, arbitrarily assigning patterns despite clear misalign-

I have no intention of stopping the use of somatostatin on patients suffering from acute pancreatitis. I consider it to be a very reasonable choice. After all, it has been standard practice in our department for many years and we've been quite satisfied with the results we've had. ***Irrelevant Authority***	
Top 1	Because many people [do/believe/support] X, X must be true/good/right/best/valid. (*Ad Populum*)
Top 2	Using [personal trait, experience, past action] as implicit proof of authority on a distinct or unrelated subject. (*Irrelevant Authority*)

Table 8: Sentence accurately classified by `o4-mini` with the 2nd ranked pattern. Although the model fails initially, the matching can still be deemed reasonable.

Teaching assistant: the homework assignment was much harder than we thought, so I think we should give a few extra points to students who completed it. Professor: that's a terrible idea. If we give everyone a perfect score for no reason, students won't bother working hard in the future. ***Extension Fallacy***	
`o4-mini`	• **Generated Form:** A proposes: "Because B, we should do X_1." Opponent replies: "If we do X_2 (an exaggerated version of X_1) for no reason, then C will happen; therefore, we must not do X_1." • **Matched Pattern:** X proposes Y; response exaggerates Y to an extreme version Y. • **Fallacy:** Extension Fallacy
`gpt-4.1-mini`	• **Generated Form:** A says B was harder than expected, so A suggests giving extra points to those who completed it. B replies: if we give everyone a perfect score for no reason, then C won't work hard; therefore, we should not give extra points. • **Matched Pattern:** Arguing that taking a minor action will cause a chain of escalating events leading to a disastrous outcome, even though no strong evidence supports the inevitability of that chain. • **Fallacy:** Hasty Generalization
`llama-3.3-70B`	• **Generated Form:** If we do A (give everyone a perfect score for no reason), then B (students won't bother working hard in the future). • **Matched Pattern:** Presenting a coincidental or correlated event as proof of causation, often ignoring other plausible causal factors or explanations. • **Fallacy:** False Cause
`gemma-3-27b-it`	• **Generated Form:** If we do A (give extra points), then B (students won't work hard). Therefore, we should not do A. • **Matched Pattern:** If P then Q; Q is true; therefore, P is true. • **Fallacy:** Deductive Fallacy

Table 10: Comparison of outputs from four models evaluated in the **MULTISTEP** configuration on LOGIC.

ment with the extracted logical form. For example, given the argument *People nowadays only vote with their emotions instead of their brains* (an instance of *Hasty Generalization*), the model `o4-mini` first extracts the logical form *All A only do B instead of C*. The model then matches this form to the pattern *Generalizing from a small sample or single event to an entire group or population*, which correctly belongs to *Hasty Generalization*. While this produces an accurate classification, the assigned pattern does not precisely correspond to the extracted logical form. In summary, while humans naturally decompose pattern matching into multiple cognitive steps, this multi-stage process proves to be challenging for current LLMs. Models struggle to bridge the gap between abstract logical patterns and their content-dependent manifestations, often failing to identify the implicit premises and unstated logical connections that underlie the reasoning chain.

6 Experiments on Further Datasets

In order to further assess the quality of LOGIC-derived patterns, we conducted a subset of the experiments on REDDIT and ELECDEBATE using the best performing model, `o4-mini`. We tested pat-

	Reddit		ElecDebate	
	Acc.	**M-F1$_1$**	**Acc.**	**M- F1$_1$**
ZERO-SHOT	82.8	82.8	67.3	50.8
DEF	82.6	82.5	65.9	54.7
LOGICAL FORMS	84.7	84.3	70.7	59.5
PATTERNS	**84.7**	**84.5**	65.5	56.3
PATTERN MATCHING	80.9	80.8	65.5	56.7
DYNAMIC ONE-SHOT	81.9	81.6	**81.7**	**70.4**
DYNAMIC + EXP	83.8	83.6	79.1	71.3
DYNAMIC + EXP + PATTERNS	79.0	78.8	78.8	72.3
SAME-DATASET PATTERNS	83.8	83.4	74.1	64.9
SAME-DATASET PATTERNS MATCHING	84.7	84.3	74.3	63.7

Table 11: Fallacy classification performance using o4-mini on Reddit and ElecDebate. **PATTERNS** method involves using patterns generated on Logic while **SAME-DATASET PATTERNS** approach includes patterns generated on the datasets Reddit and ElecDebate themselves. Acc.: Accuracy; M-F1: Macro-F1.

terns extracted from Logic, restricted to the two datasets' classes (first eight rows in Table 11) and patterns extracted from the datasets themselves (latest two rows in Table 11). Reddit patterns show a prevalence of linguistic markers over logical forms while ElecDebate ones emphasize stronger logical formalization, incorporating symbolic formalism.

Consistent with previous findings, logical pattern incorporation outperforms competing approaches on Reddit. Moreover, Logic-based and Reddit-based patterns yield comparable results. While taxonomy alignment prevents direct comparison, results from supervised and unsupervised methods (Sahai et al., 2021; Lei and Huang, 2024; Pan et al., 2024; Yeh et al., 2024) are consistent with our findings (see Appendix C). Only a comparison with Lei and Huang (2024) (Macro F_1=81.3%) is possible: **PATTERNS** and **SAME-DATASET PATTERNS** outperform their results.

Regarding ElecDebate, Table 11 shows that **DYNAMIC ONE-SHOT** yields the best performance, possibly due to the predominant presence of the class *Emotional Language* (62.5% of test set) whose detection may particularly benefit from similar worded examples. Indeed, **SAME-DATASET PATTERNS** achieve competitive results with respect to Goffredo et al. (2023); Pan et al. (2024) (see Appendix C). These experiments showed a fair generalization of Logic-derived patterns on other datasets, with the additional advantage of not requiring labeled data to re-extract the patterns.

In order to prove the broader applicability of our approach beyond Logic-specific patterns, we tested patterns generated from the other two datasets on Logic.

Table 12 demonstrates solid results, validating findings on Logic and proving the robustness and transferability of our pattern-based methodology. Notice that accuracy is not directly comparable with values in Table 5 since Reddit and ElecDebate have a subset of the classes of Logic.

	$\text{Logic}_{\text{Reddit}}$	$\text{Logic}_{\text{ElecDebate}}$
	Acc.	**Acc.**
SAME-DATASET PATTERNS	90.8	87.1
SAME-DATASET PATTERNS MATCHING	88.3	87.5
LOGIC PATTERNS	89.1	83.7
LOGIC PATTERNS MATCHING	90.0	87.1

Table 12: Fallacy classification performance using o4-mini on Logic. Logic_X refers to Logic restricted to the classes from dataset X. **SAME-DATASET PATTERNS** approach includes patterns generated on non-Logic dataset X while **LOGIC PATTERNS** involves using Logic-derived patterns restricted to the classes of dataset X.

7 Conclusions

Fallacy detection is a challenging yet critical task to solve. Since fallacies often manifest in nuanced and context-dependent forms, purely abstract representations are insufficient to characterize the full spectrum of ways a fallacy can appear in natural language, thus motivating the need to combine logical structure with context-level linguistic cues. We present an experimental framework that inductively extracts context-aware structural patterns from fallacious arguments and their explanations, demonstrating that incorporating such patterns significantly enhances fallacy classification performance. Specifically, pattern-based classification achieves 73.5% accuracy on Logic, significantly outperforming prior unsupervised approaches, and 74.2% including one-shot examples. Being data-driven, these patterns are not bound to a fixed set of fallacies and can flexibly capture the diverse nuances through which each fallacy type manifests. Notably, reasoning models demonstrate consistently superior performance across all experimental configurations. Moreover, experiments on additional datasets confirm that the extracted patterns generalize effectively across domains, establishing data-driven pattern extraction as an effective method to generate valid and generalizable logical representations.

8 Limitations

While this work demonstrates the efficacy of large language models in detecting logical fallacies by exploiting the underlying logical structure of sentences, it has several limitations. First, we intentionally generated patterns exclusively from the LOGIC dataset due to the quality and straightforward structure of its sentences. We are aware, however, that it does not fully cover the complex and multi-faceted spectrum of fallacies. Furthermore, our work is based on a small sample of LLMs. Nevertheless, we selected a diverse and representative subset, including models from different providers, with varying sizes and reasoning capabilities.

9 Ethics Statement

Logical fallacies can reinforce societal bias and facilitate the spread of misinformation, leading to harmful consequences for society. This work focuses on leveraging LLMs for detecting logical fallacies in argumentation and should not be employed to manipulate discourse by exploiting identified reasoning patterns. Furthermore, this approach risks amplifying existing LLM biases, potentially causing unfair detection. We acknowledge these limitations and encourage future bias mitigation research. We are aware of the environmental impact of large-scale LLMs usage. However, this study exclusively employs inference-only methods, significantly reducing computational requirements compared to training approaches. All datasets are used in accordance with their license and they have been checked for personally identifying and offensive content.

Acknowledgements

This publication is part of the project PNRR-NGEU, which has received funding from the MUR - DM 629/2024. We would like to thank the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI), for also funding this work by grant NPRP14C0916-210015.

References

Tariq Alhindi, Smaranda Muresan, and Preslav Nakov. 2024. Large language models are few-shot training example generators: A case study in fallacy recognition.

John B. Bacon, Michael Detlefsen, and David Charles McCarty. 1999. *Logic from A to Z: The Routledge Encyclopedia of Philosophy Glossary of Logical and Mathematical Terms*. Routledge.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners.

Irving Marmer Copi, Carl Cohen, and Kenneth McMahon. 1953. *Introduction to Logic*. Macmillan, New York, NY, USA.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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 5636–5646, Hong Kong, China. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

H.J. Gensler. 2010. *The to Z of Logic*. Number v. 169 in G - Reference,Information and Interdisciplinary Subjects Series. Bloomsbury Academic.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Ana Gutiérrez-Mandingorra, Stella Heras, and Javier Palanca. 2024. Detecting disinformation through computational argumentation techniques and large language models. In *Proceedings of the 24th Workshop on Computational Models of Natural Argument (CMNA 2024)*, volume 3769 of *CEUR Workshop Proceedings*, pages 46–51.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

C.L. Hamblin. 1970. *Fallacies*. University paperbacks. Methuen.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Representation learning on graphs: Methods and applications.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. 2024. A closer look at the self-verification abilities of large language models in logical reasoning.

Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. Large language models are better logical fallacy reasoners with counterargument, explanation, and goal-aware prompt formulation.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ralph Henry Johnson and J. Anthony Blair. 1977. *Logical Self-Defense*. Toronto, Canada.

Yuanyuan Lei and Ruihong Huang. 2024. Boosting logical fallacy reasoning in LLMs via logical structure tree.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding.

Gionnieve Lim and Simon T. Perrault. 2024. Evaluation of an llm in identifying logical fallacies: A call for rigor when adopting llms in hci research.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2025. Openai o3 and o4-mini system card.

Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. 2024. Are LLMs good zero-shot fallacy classifiers?

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*. Association for Computational Linguistics.

Paul Reisert, Naoya Inoue, Tatsuki Kuribayashi, and Kentaro Inui. 2018. Feasible annotation scheme for capturing policy argument reasoning using argument templates. In *Proceedings of the 5th Workshop on Argument Mining*, pages 79–89, Brussels, Belgium. Association for Computational Linguistics.

Irfan Robbani, Paul Reisert, Surawat Pothong, Naoya Inoue, Camélia Guerraoui, Wenzhi Wang, Shoichi Naito, Jungmin Choi, and Kentaro Inui. 2024. Flee the flaw: Annotating the underlying logic of fallacious arguments through templates and slot-filling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20524–20540, Miami, Florida, USA. Association for Computational Linguistics.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore. Association for Computational Linguistics.

Ramon Ruiz-Dolz and John Lawrence. 2025. An explainable framework for misinformation identification via critical question answering. *Preprint*, arXiv:2503.14626.

Devendra Sachan, Yuhao Zhang, Peng Qi, and William L. Hamilton. 2021. Do syntax trees help pre-trained transformers extract information?

Saumya Sahai, Oana Balalau, and Roxana Horincar. 2021. Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 644–657, Online. Association for Computational Linguistics.

Zhivar Sourati, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2023a. Case-based reasoning with language models for classification of logical fallacies.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông-Ân Sandlin, and Alain Mermoud. 2023b. Robust and explainable identification of logical fallacies in natural language arguments.

Thomas Storer. 1949. Carl g. hempel and paul oppenheim. studies in the logic of explanation. philosophy of science, vol. 15 (1948), pp. 135–175. *Journal of Symbolic Logic*, 14(2):133–133.

Nicole Teo, Donghao Huang, Erik Cambria, and Zhaoxia Wang. 2025. Large language models for logical fallacy detection. In *Trends and Applications in Knowledge Discovery and Data Mining*, pages 387–398, Singapore. Springer Nature Singapore.

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. TWEETSPIN: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448, Seattle, United States. Association for Computational Linguistics.

Douglas N. Walton. 2008. *Informal logic : a pragmatic approach*, second edition. edition. Cambridge University Press, Cambridge.

Xiaoou Wang, Elena Cabrio, and Serena Villata. 2025. When automated fact-checking meets argumentation: Unveiling fake news through argumentative evidence. *Argument & Computation*, 16(3):405–424.

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.

Zihao Xu, Junchen Ding, Yiling Lou, Kun Zhang, Dong Gong, and Yuekang Li. 2026. Socrates or smartypants: Testing logic reasoning capabilities of large language models with logic programming-based test oracles.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training.

A Implementation Details

In experiments where the task consisted of returning only the fallacy label, we set the temperature to 0, with the exception of `o4-mini`, `gpt-4o` and `deepseek-r1`. In all other experiments, the standard configuration was kept. Multiple prompt configurations were evaluated for each approach.

B Fallacy Datasets

B.1 Logic

The dataset LOGIC (Jin et al., 2022) contains the following 13 fallacy classes: *Faulty Generalization (Hasty Generalization)*, *Ad Hominem*, *Ad Populum*, *Circular Claim (Circular Reasoning)*, *False Cause (False Causality)*, *Appeal to Emotion (Emotional Language)*, *Fallacy of Relevance (Red Herring)*, *Deductive Fallacy*, *Intentional Fallacy*, *Fallacy of Extension (Extension Fallacy)*, *False Dilemma (Black-and-White Fallacy)*, *Fallacy of Credibility (Irrelevant Authority)* and *Equivocation*. The names in the parentheses are the actual names used in our experiments.

REDDIT	ELECDEBATE
• Ad Populum	• Ad Hominem
• Irrelevant Authority	• Irrelevant Authority
• Hasty Generalization	• Emotional Language
• Slippery Slope	• Slippery Slope
• Black-and-White Fallacy	• False Cause

Table 13: Fallacy classes in REDDIT and ELECDEBATE used in our experiments.

B.2 Reddit

The dataset REDDIT (Sahai et al., 2021) contains 8 fallacy classes: *Appeal to Authority (Irrelevant Authority)*, *Appeal to Majority (Ad Populum)*, *Appeal to Nature*, *Appeal to Tradition*, *Appeal to Worse Problems*, *Black-and-White fallacy*, *Hasty Generalization* and *Slippery Slope*. It contains the class *No Fallacy* as well. The names in parentheses are the actual labels used. In our experiments, only the classes included in LOGIC are retained (Table 13). We can keep the class *Slippery Slope* because two generated patterns for *Hasty Generalization* correspond to it.

B.3 ElecDebate

The dataset ELECDEBATE (Goffredo et al., 2023) contains the following 6 fallacy classes: *Ad Hominem*, *Appeal to Emotion (Emotional Language)*, *Appeal to Authority (Irrelevant Authority)*, *Slippery Slope*, *False Cause* and *Slogan*. The names in parentheses are the actual labels used. In our experiments, only the classes included in LOGIC are retained (Table 13).

C Baselines

We consider only the classes of REDDIT and ELECDEBATE in common to LOGIC. For this reason, direct comparison with prior work is generally not possible. However, for REDDIT, Lei and Huang (2024) provide classwise F_1 scores, allowing us to compute Macro F_1 and compare our results. Tables 14 and 15 present the comparison with prior work for both datasets.

D Additional Experiments

We are going to report some other experimental setups that have been explored, including some basic baselines that we have not included in Section 5.

Method	Macro F_1
Supervised	
Sahai et al. (2021)	58.4
Lei and Huang (2024)†	81.3
Pan et al. (2024)	83.2
Unsupervised	
Pan et al. (2024)	81.1
Yeh et al. (2024)	81.0
Ours	
PATTERNS	**84.5**
SAME-DATASET PATTERN MATCHING	**84.3**

Table 14: Performance comparison on REDDIT.† indicates that Macro F_1 is computed on the exact same classes as LOGIC.

Method	Macro F_1
Supervised	
Goffredo et al. (2023)	**73.9**
Pan et al. (2024)	62.3
Unsupervised	
Pan et al. (2024)	44.5
Ours	
SAME-DATASET PATTERN	64.9
DYNAMIC ONE-SHOT	70.4

Table 15: Performance comparison on ELECDEBATE.

D.1 Prompt design

- **EXP**: to investigate whether explicit reasoning improves performance, we implemented a baseline that not only provides fallacy names but also requests the model to generate a two-sentence explanation for its classification decision, testing whether forcing the model to articulate its reasoning leads to better outcomes. The two-sentence constraint was intentionally designed to keep explanations concise and manageable for manual inspection of explanations.

- **GUIDELINES**: to leverage the model's classification errors for improvement, we develop guidelines derived from observed mistakes. We conduct pattern matching evaluation on the validation set and collect misclassified instances. For each class, we provide the model with incorrectly classified examples and prompt it to generate comprehensive detection guidelines (as can be seen from table 16), given our generated pattern as a reference. These guidelines are then adopted to evaluate the test set. Notably, only guidelines produced by `o4-mini` and partially by `gpt-4.1-mini` incorporate a little structural and logical information such as common connectors or logical forms while the majority of guidelines content across models focuses primarily on semantic characteristics rather than structural patterns.

Fallacy	Irrelevant Authority
Core definition	A fallacy that treats an individual's status, title, or popularity as proof of a claim when their expertise or relevance to the topic is absent or insufficient.
Key indicators	Argument rests on "X says so" without independent support. Authority cited has no recognized expertise in the claim's domain. No substantive evidence beyond the authority's endorsement.
Typical confusion patterns	Ad Populum: group popularity vs. single authority endorsement. Appeal to Tradition: "has always been done by experts" vs. citing irrelevant experts. Equivocation: shifting word senses vs. relying on irrelevant credentials.

Table 16: Guidelines relative to the *Irrelevant Authority* fallacy generated by `o4-mini`.

Model	EXP		GUIDELINES	
	Acc.	F_1	Acc.	F_1
`o4-mini`	61.3	61.5	65.5	65.7
`gpt-4o`	60.4	53.3	62.6	56.6
`deepseek-r1`	61.8	54.8	63.5	57.9
`gpt-4.1-mini`	57.5	57.9	60.5	60.6
`llama-3.3-70B`	56.1	56.5	52.8	53.3
`gemma-3-27b-it`	59.1	60.8	58.8	59.4

Table 17: Logical fallacy classification performance on additional experiments. F_1 denotes Macro-F_1.

D.2 Results

EXP's (Table 17) results show that requesting the model to articulate the reasoning does not really cause any improvement. Specifically, certain classes such as *Intentional Fallacy* and *Extension Fallacy* exhibit extremely low F_1 scores under the non-reasoning models (0.027 and 0.13 respectively on average), indicating performance

Text	Explanation	Gold
The Bible is true because God exists, and God exists because the Bible says so.	The argument uses its conclusion as a premise, creating a logical loop without independent evidence. ***Circular Reasoning***	***Circular Reasoning***
My friend said that if you sneeze more than three times, you have the corona virus.	The argument assumes sneezing three times indicates the virus, generalizing a symptom without considering other causes. ***Hasty Generalization***	***Irrelevant Authority***

Table 18: Examples from `GPT-4.1-mini` in the **EXP** setting: the first is correctly classified; the second is misclassified because the explanation, while coherent, fails to capture the underlying fallacy.

deterioration compared to the **ZERO-SHOT** baseline. This proves that models process surface-level semantic patterns without being able to access the multi-layered intentional structures behind reasoning (Table 18).

Including **GUIDELINES** yields only modest results. While these guidelines are designed to provide comprehensive fallacy knowledge, they appear to lack the appropriate type of information from which models can benefit. Indeed, providing explicit information about the underlying logical structure proves significantly more beneficial for model performance.

E Syntax-augmented roBERTa

Sachan et al. (2021) introduces a syntax-augmented model that incorporates dependency tree information into pre-trained BERT-based (Devlin et al., 2019) transformers through specialized Graph Neural Networks (GNNs) (Hamilton et al., 2017) that process dependency trees. The authors introduce two distinct fusion strategies to integrate syntactic structure into BERT representation. We adopted specifically roBERTa-large (Zhuang et al., 2021) in the attempt to perform a syntax-driven examples selection. Further details about the implementation are available in Sachan et al. (2021).

A Three-Level Audit of LLM Alignment for Argument Quality Assessment

Wei-Fan Chen*[1,2,3] **Jinming Yu***[2] **Lucie Flek**[2,3]
[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2]Bonn-Aachen International Center for Information Technology, University of Bonn, Germany
[3]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
wf.chen@aist.go.jp, s6jiyuuu@uni-bonn.de, flek@bit.uni-bonn.de

Abstract

Large Language Models (LLMs) are increasingly used as automated evaluators of argument quality. However, existing studies typically assess models only through their agreement with human scores, leaving the reasoning process behind these judgments unexplored. In this paper, we propose a three-level audit framework for evaluating the reliability of LLM-based argument quality assessment. The framework distinguishes between (1) **surface alignment**, measuring agreement between LLM-predicted scores and human annotations; (2) **instructional alignment**, assessing whether generated rationales adhere to the intended evaluation criteria; and (3) **faithfulness alignment**, examining whether predicted scores are supported by the generated rationales. To operationalize this audit, we introduce structural rationale prompting,[1] which guides LLMs to generate structured justifications before assigning scores across 11 dimensions of the Dagstuhl-15512 argument quality corpus. We evaluate several LLMs under this framework and find that structural rationale prompting substantially improves agreement with human annotations compared to definition-based prompting. Furthermore, the generated rationales generally follow the evaluation instructions and remain highly consistent with the predicted scores. Overall, our results suggest that auditing LLM evaluators beyond surface score agreement provides deeper insight into the reliability and transparency of LLM-based argument quality assessment.

1 Introduction

The assessment of argument quality remains a foundational challenge in argument mining (Habernal and Gurevych, 2016; Wachsmuth et al., 2017; Lawrence and Reed, 2019). While the theoretical roots of argumentation trace back to Ancient Greece (Aristotle, 2007), operationalizing these concepts in computational settings requires taxonomies that decompose argument quality into multiple measurable dimensions (Blair, 2011; Wachsmuth et al., 2017). Annotating these dimensions, however, is a labor-intensive and time-consuming process that often requires specially trained annotators and carefully designed guidelines (Toledo et al., 2019; Lauscher et al., 2020). As a result, expert-labeled datasets remain limited, creating a significant bottleneck for training automatic argument quality evaluators and for advancing downstream tasks in argument mining.

Recent advances in Large Language Models (LLMs) have prompted researchers to investigate whether such models can follow natural language instructions to assist with data annotation and evaluation tasks (Gilardi et al., 2023; Chiang and Lee, 2023; Mirzakhmedova et al., 2024). Experimental results across several studies report substantial agreement between LLM judges and human judges, suggesting that LLMs may serve as scalable evaluators in various NLP settings (Zheng et al., 2023; Kocmi and Federmann, 2023). However, these models are typically employed as *black-box evaluators*, whose reliability is assessed primarily through agreement with human scores. Even when reasoning-oriented prompting strategies such as Chain-of-Thought prompting (Wei et al., 2022) are applied, it often remains unclear whether the generated explanations follow the intended evaluation criteria or meaningfully justify the final scores produced by the model. Consequently, the reasoning processes underlying LLM-based evaluations remain largely under-examined.

In this work, we argue that evaluating LLM-based argument quality assessment requires auditing not only predicted scores but also the reasoning processes that lead to them. In practice, an LLM evaluator may fail in several ways: its predicted scores may diverge from human judg-

[1]All the codes and prompts used in the paper are available at https://github.com/aist-cwf/argmining26-argument-quality-assessment.

*Equal Contribution.

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 26–38
July 3, 2026 ©2026 Association for Computational Linguistics

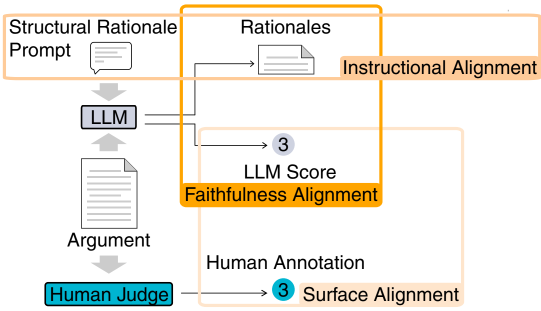


Figure 1: Overview of the proposed three-level audit framework for LLM-based argument quality assessment. Given an argument and an evaluation prompt, the LLM produces rationales and a predicted score. Surface alignment measures agreement between the LLM score and human annotations. Instructional alignment evaluates whether the generated rationales follow the evaluation criteria. Faithfulness alignment examines whether the predicted score is supported by the generated rationales.

ments, its explanations may not follow the intended evaluation criteria, or its reasoning may not actually support the predicted scores. To systematically analyze these potential failure modes, we propose a three-level audit framework for LLM-based argument quality assessment. As illustrated in Figure 1, the framework distinguishes three complementary alignment dimensions: surface alignment, which measures agreement between LLM-predicted scores and expert human annotations; instructional alignment, which evaluates whether generated rationales adhere to the intended evaluation instructions; and faithfulness alignment, which examines whether predicted scores are supported by the generated rationales. These dimensions guide our investigation of the reliability and transparency of LLM-based argument evaluation through the following research questions.

(1) Surface Alignment: To what extent do LLM scores correlate with human annotations across diverse argument quality dimensions?
(2) Instructional Alignment: To what extent do the rationales generated by LLMs adhere to the prompted instructions and criteria defined for each argument quality dimension?
(3) Faithfulness Alignment: To what extent are the predicted scores supported by the generated rationales?

To operationalize this audit, we introduce structural rationale prompting, a prompting strategy that guides LLMs to generate step-wise rationales based on predefined evaluation criteria before assigning quality scores. We apply this approach to assess argument quality across all 11 dimensions of the Dagstuhl-15512 argument quality corpus (Wachsmuth et al., 2017). Using the generated rationales and predicted scores, we analyze the alignment between LLM outputs and human judgments across the three alignment dimensions defined above.

The first research question examines whether LLMs can reliably annotate argument quality in a manner comparable to human experts. To address this question, we evaluate the statistical alignment between LLM-predicted scores and expert human annotations, measuring agreement using Krippendorff's α and prediction error using RMSE. We further analyze whether generating structured rationales improves agreement between LLMs and human experts. Among the evaluated models, `mistral-small-latest` achieves the strongest performance (average $\alpha = 0.80$ and average RMSE $= 0.63$). Moreover, structural rationale prompting yields substantially higher agreement ($\alpha = 0.67$ vs. 0.28, both using GPT-3.5-Turbo) than the "expert prompt" used in Mirzakhmedova et al. (2024), which provides only definitions of the argument quality dimensions.

While the first research question evaluates alignment between LLM-predicted scores and human annotations, the second examines the quality of the generated rationales. Because the Dagstuhl-15512 dataset does not provide gold-standard justifications, we assess whether LLM-generated rationales adhere to the structured evaluation instructions. We validate an independent judge LLM (GPT-5.2) against human expert annotations on a small subset, yielding a Krippendorff's α of 0.56, and then use it to evaluate the remaining rationales. On a 3-point scale, the results show consistently high levels of instructional alignment across the four evaluated models, with scores ranging from 2.00 (`Mistral-7B-Instruct`) to 2.79 (`Mistral-small-latest`).

The third research question investigates whether the models' predicted scores are supported by their stated rationales. To assess this relationship, annotators are asked to infer argument quality scores based solely on the generated rationales, without access to the original model predictions. These rationale-based scores are then compared with the model-predicted scores to derive a measure of faithfulness. As in the previous evaluation, we validate an independent judge LLM (GPT-5.2) against human expert annotations on a small subset (Krippendorff's $\alpha = 0.56$) before applying it to the remaining instances. On a 3-point scale, the results indicate consistently high faithfulness alignment across all evaluated models, with scores ranging from 2.80 (`Mistral-7B-Instruct` and `GPT-3.5-Turbo`) to 2.96 (`Llama-3.3-70B-Instruct`).

Our contributions are as follows: (1) We introduce a reproducible auditing framework for analyzing the reliability of LLM-based argument quality evaluation through three alignment dimensions: surface alignment, instructional alignment, and faithfulness alignment. (2) We propose structural rationale prompting, a prompting strategy that guides LLMs to generate structured justifications based on predefined argument quality criteria before assigning scores. (3) Through experiments on the Dagstuhl-15512 argument quality corpus, we show that structural rationale prompting improves alignment with human annotations and that modern LLMs generate rationales that largely follow evaluation instructions while remaining highly consistent with their predicted scores.

2 Related Work

This section first reviews the evolution of argument quality assessment from theory-grounded taxonomies to neural scoring models. We then examine the emerging literature on LLM Interpretability, focusing on generating faithful rationales within structured reasoning frameworks.

Argument Quality Assessment The computational argument quality assessment has evolved across three distinct paradigms: theory-driven feature engineering, neural representation learning, and the current era of generative modeling.

Early work in argument quality assessment was heavily grounded in classical rhetoric and informal logic. Following the Aristotelian tradition, researchers initially focused on identifying the components of an argument using the Toulmin model (Toulmin, 2003) or Walton's argumentation schemes (Walton et al., 2008). Habernal and Gurevych (2016) pioneered the *convincingnes task*, using crowdsourcing to determine which of two arguments is more persuasive. This was further refined by Wachsmuth et al. (2017), who established the foundational taxonomy of 11 fine-grained quality dimensions, providing the Dagstuhl-15512 corpus which remains the gold standard for multi-dimensional assessment.

With the rise of deep learning, the field shifted toward neural models that could capture latent linguistic features. Potash et al. (2017) and Stab and Gurevych (2017) explored the use of Recurrent Neural Networks for scoring argument persuasiveness and structure. The introduction of Transformer-based models like BERT (Devlin et al., 2019) significantly improved performance on quality scoring tasks, as seen in the work of Gretz et al. (2020), who released a large-scale dataset of 30k arguments scored for quality. Subsequent research explored cross-domain generalizability (Toledo et al., 2019) and the use of graph neural networks to model the relational structure between conflicting arguments (Ye and Teufel, 2021).

Most recently, the paradigm has shifted toward using LLMs as judges. For example, Stahl et al. (2025) fine-tuned an LLM via the proposed specialized instruction in various computational argument tasks including quality assessment. In fact, while models like GPT-4 have demonstrated high correlation with human scores in general text evaluation (Chiang and Lee, 2023), their application to the specific nuances of argumentation is an ongoing area of research. Liu et al. (2023) introduced G-Eval, which uses Chain-of-Thought to improve evaluation consistency. However, as noted by Wang et al. (2023), LLM evaluators are susceptible to self-preference bias and hallucinated justifications. Our work builds on this line of inquiry by moving beyond simple scoring to audit the faithfulness of the rationales generated.

Rationale Generation and LLM Faithfulness The emergence of Large Language Models has shifted focus from purely discriminative tasks to generative reasoning. The introduction of Chain-of-Thought prompting by Wei et al. (2022) demonstrated that forcing a model to generate intermeidary reasoning steps significantly enhances performance on complex tasks. In the domain of evaluation, Liu et al. (2023) and Chan et al. (2023) found that eliciting these rationales leads to higher alignment with human-assigned scores. However, the nature of these rationales remains a subject of intense scrutiny. Researchers have debated whether these outputs represent a process-oriented logic or merely a result-oriented justification (Zelikman et al., 2022; Lightman et al., 2023).

A critical concern in LLM interpretability is faithfulness: the degree to which a generated

Quality Dimension	Definition
Cogency	
Local acceptability	Premises worthy of being believed.
Local relevance	Premises accept/reject conclusion.
Local sufficiency	Premises enough to draw conclusion.
Effectiveness	
Credibility	Makes author worthy of credence.
Emotion appeal	Makes audience open to arguments.
Clarity	Uses correct, unambiguous language and avoids deviation from the issue.
Appropriateness	Language supports credibility and emotions, in proportion to the issue.
Arrangement	Argues in the right order.
Reasonableness	
Global acceptability	Audience accepts use of argument.
Global relevance	Argument helps arrive at conclusion.
Global sufficiency	Enough rebuts counterarguments.

Table 1: Definition and categorization of each quality dimension, adapted from Wachsmuth et al. (2017) and Mirzakhmedova et al. (2024).

Quality Dimension	#Instances
Cogency	
Local acceptability	82
Local relevance	99
Local sufficiency	113
Effectiveness	
Credibility	115
Emotion appeal	130
Clarity	89
Appropriateness	53
Arrangement	81
Reasonableness	
Global acceptability	96
Global relevance	66
Global sufficiency	136

Table 2: The number of instances in each dimension where there are perfect agreement among the annotators.

rationale accurately reflects the model's internal decision-making process. Jacovi and Goldberg (2020) established the formal criteria for faithfulness, distinguishing it from mere plausibility (how human-like an explanation sounds). Chen et al. (2025) examined faithfullness hallucination by creating a heatmap-like "Halumap," and Chen et al. (2024) tried to generate more faithful outputs based on a multi-task learning approach. Recent audits have revealed a disturbing trend of post-hoc rationalization, where models generate convincing but fundamentally disconnected explanations for their outputs (Turpin et al., 2023).

To move beyond black-box behavior, several studies have proposed methods to audit the causal link between rationales and outputs. Lyu et al. (2023) explored the use of faithful chain-of-thoughts by enforcing symbolic intermediate steps. Meanwhile, the use of Judge LLMs to evaluate the quality and logic of another model's reasoning has become a standard diagnostic tool (Zheng et al., 2023). Our study contributes to this line of work by applying these "faithfulness audits" to the specific, highly subjective domain of 11-dimension argumentation quality.

3 Experimental Setup

This section presents the experimental setup that underlies the three alignment research questions investigated in this study. We begin by introducing the argument quality assessment task, followed by descriptions of the dataset, the prompting strategies used to elicit model judgments and rationales, and the LLMs evaluated in our experiments.

3.1 Task

Following the argument quality taxonomy proposed by Wachsmuth et al. (2017), our task is to assess arguments along 11 fine-grained dimensions of argument quality. These dimensions capture different aspects of argumentative strength and effectiveness. Table 1 summarizes the dimensions together with their definitions.

3.2 Dataset

Based on this taxonomy, Wachsmuth et al. (2017) constructed the Dagstuhl-15512 argument quality corpus by annotating 320 arguments drawn from the UKPConvArgRank dataset (Habernal and Gurevych, 2016). Each argument was evaluated by three expert annotators with respect to the defined quality dimensions. The reported interannotator agreement, measured using Krippendorff's α, ranges from 0.23 to 0.60 across dimensions. Annotators assigned scores on a 3-point Likert scale (1 = low, 2 = medium, 3 = high) or selected *cannot judge*, which occurred only once among approximately 14,000 individual assessments.

To obtain a single reference label for comparison with LLM predictions, we aggregate the scores from the three annotators using MACE (Hovy et al., 2013).[2] MACE estimates the reliability of each annotator and infers a latent true label based on these reliability estimates, thereby providing a principled aggregation of multiple annotations.

We also follow Mirzakhmedova et al. (2024) in

[2]We also experimented with simple score averaging and observed very similar results.

Task: Evaluate the quality of the Local Sufficiency of the given argument in the context of a debate forum. The input includes an issue, a stance, and a corresponding argument. Your goal is to think step-by-step about the structural flow of the argumentation and assign a quality score for **Local Sufficiency** (from 1 to 3).

Input
- **Issue**: `ISSUE`
- **Stance**: `STANCE`
- **Argument**: `ARGUMENT`

Definition of Local Sufficiency: The premises of an argument should be seen as sufficient if, together, they provide enough support to make it rational to draw the argument's conclusion. If you identify more than one conclusion in the comment, try to adequately weight the sufficiency of the premises for each conclusion when judging about their "aggregate" sufficiency—unless there are particular premises or conclusions that dominate your view of the author's argumentation. Notice that you may see premises as sufficient even though you do not personally accept all of them, i.e., sufficiency does not presuppose acceptability.

Step-by-Step Instructions:

(1) Identify the conclusion — Determine the main claim that the author is trying to prove or support.
(2) List the premises — Extract the key reasons or pieces of evidence offered to support the conclusion. Ensure they are clearly separable and relevant.
(3) Evaluate logical connections — Analyze whether the premises logically and directly support the conclusion without major gaps or irrelevant claims.
(4) Assess premise quality and evidence support — Check if the premises are factually sound, supported by evidence, and free from bias or unsupported assumptions.
(5) Make a global judgment
  - Score 3 (High) — Premises are relevant, coherent, and collectively sufficient to justify the conclusion.
  - Score 2 (Medium) — Premises are partially sufficient but lack depth or contain minor gaps in reasoning.
  - Score 1 (Low) — Premises are weak, biased, or fail to support the conclusion logically.

Examples:
- Argument: *example_argument*
- Issue: *example_issue*
- Stance: *example_stance*
- Reasoning: *example_reasoning*
- Final score: *example_score*

Table 3: The structural rationale prompt for assessing the *local sufficiency* quality dimension. The `teletype` tokens are the placeholders for the argument to be evaluated. Two examples for few shots learning are provided in the end but are omitted here due to space constraints. The full example and the model outputs can be found in Sec A.1.

constructing a *perfect-agreement subset*, consisting of instances for which all annotators assigned the same score. This subset represents cases with unambiguous human judgments and therefore provides a useful setting for analyzing model behavior under high annotation certainty. The number of instances in the perfect-agreement subset is reported in Table 2.

3.3 Prompt Design

Prompt design for LLMs is known to be highly sensitive, as even small variations in wording can lead to substantially different outputs. Rather than focusing on prompt engineering for optimal performance, this work introduces a structural rationale prompting framework designed to guide language models through a structured reasoning process before assigning argument quality scores. The goal of this framework is to encourage models to approximate the reasoning procedure of expert annotators when evaluating arguments.

Structural Rationale Prompt Table 3 illustrates the structural rationale prompt using the *local sufficiency* dimension as an example. The prompt first provides the `issue`, `stance`, and `argument`, followed by the definition of local sufficiency taken from the annotation guidelines in Wachsmuth et al. (2017). Based on this definition and prior computational approaches to modeling the dimension, we formulate step-by-step instructions that guide the model in producing a structured rationale prior to assigning a score.

For the local sufficiency dimension, both the definition and prior computational work (Rahimi et al., 2014) suggest that evaluation requires several reasoning steps: (1) identifying the premises and the conclusion of the argument, (2) assessing the logical support provided by each premise for the conclusion, and (3) aggregating these assessments to determine whether the premises are collectively sufficient. Based on this reasoning procedure, we formulate the structured rational prompt shown in the example prompt. Importantly, our goal is not to claim that these prompts represent an optimal design for argument quality annotation. Instead, they serve as an analytical tool for examining whether

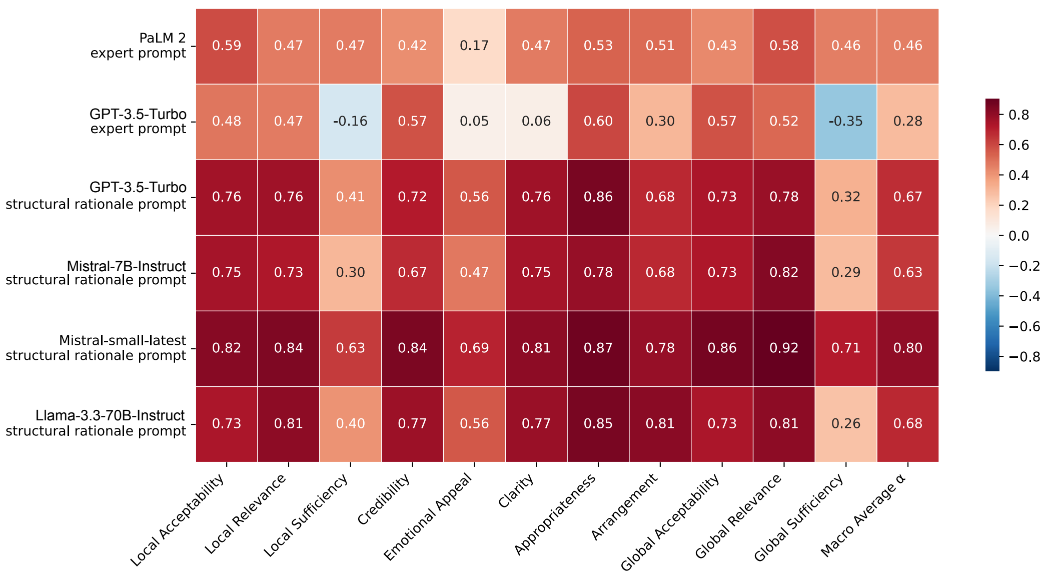


Figure 2: Krippendorff's α measuring agreement between human annotations and LLM predictions across models and prompting strategies. Results for the expert prompt are taken from Mirzakhmedova et al. (2024). Structural rationale prompting consistently improves agreement across the evaluated dimensions.

LLMs follow the intended evaluation procedure and whether their predicted scores are supported by the generated rationales.

All experiments are conducted three times, and we report the average performance across runs. The few-shot examples used in the prompts are sampled from the Dagstuhl-15512 corpus and are excluded from the corresponding evaluation sets to avoid overlap between demonstration and test instances.

Expert Prompt To enable comparison with the results reported by Mirzakhmedova et al. (2024), we also evaluate the so-called "expert prompt." This prompting strategy provides the LLM with only the definition of each argument quality dimension taken from the annotation guidelines, without additional step-by-step instructions.

3.4 Models

As our evaluation setup, we prompt several LLMs with varying architectures and model scales: (1) Llama-3.3-70B-Instruct, (2) GPT-3.5-Turbo, a large proprietary model, (3) Mistral-7B-Instruct, and (4) Mistral-small-latest (approximately 24B parameters). These models cover both open-weight and commercial systems and represent a range of model capacities. To ensure comparability across models, we use the same prompts and experimental settings ($temperature = 0.7$, and $top\_p = 0.9$) for all models, allowing us to better isolate the effects of model scale and architecture on argument quality assessment.

4 Surface Alignment

The first research question examines surface alignment, which measures **the degree of agreement between LLM-predicted argument quality scores and human annotations**. Our goal is to determine whether structural rationale prompting improves annotation performance compared with the prompting strategy used in prior work. To this end, we compare our model predictions with the results reported by Mirzakhmedova et al. (2024) on the perfect-agreement subset, where all human annotators assigned the same score and the ground truth is therefore unambiguous. Figure 2 reports Krippendorff's α between human annotations and LLM predictions across models and prompting strategies. Overall, structural rationale prompting leads to consistently higher agreement across the evaluated dimensions. Notably, despite having fewer parameters than GPT-3.5-Turbo and Llama-3.3-70B-Instruct, Mistral-small-latest achieves the strongest overall agreement.

To complement the agreement analysis, Table 4 reports the RMSE of the four models with respect to the aggregated reference labels on the whole dataset. The results largely corroborate the findings from Krippendorff's α: Mistral-small-latest achieves the lowest overall prediction error across the evaluated dimensions. Among the remaining

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Llama-3.3-70B-Instruct	0.74	0.73	**0.71**	**0.64**	0.76	0.73	0.73	0.76	**0.61**	0.71	0.77	0.72
GPT-3.5-Turbo	0.78	0.83	0.91	0.77	0.79	0.80	0.79	0.86	0.73	0.89	0.66	0.80
Mistral-7B-Instruct	**0.70**	0.88	0.82	0.80	0.85	0.75	0.72	0.87	0.78	0.79	0.67	0.78
Mistral-small-latest 2-shots	0.73	**0.71**	**0.71**	0.66	**0.70**	**0.67**	**0.67**	**0.74**	0.66	**0.68**	**0.58**	**0.68**
Mistral-small-latest 0-shots	0.85	0.79	0.74	0.69	0.79	0.69	0.75	0.84	0.67	0.72	0.80	0.74

Table 4: RMSE of the four models and `mistral-small-latest` under zero-shot setting. The abbreviations denote the following dimensions: **L**ocal **A**cceptability, **L**ocal **R**elevance, **L**ocal **S**ufficiency, **Cr**edibility, **Em**otional Appeal, **Cl**arity, **Ap**propriateness, **Ar**rangement, **G**lobal **A**cceptability, **G**lobal **R**elevance, and **G**lobal **S**ufficiency. The best-performing setting for each dimension is highlighted in bold.

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Llama-3.3-70B-Instruct	2.54	2.93	2.98	**3.00**	2.40	**3.00**	**3.00**	2.96	**2.46**	2.94	2.17	2.76
GPT-3.5-Turbo	1.98	1.93	2.00	2.14	2.25	1.99	2.32	2.77	2.29	2.00	2.25	2.17
Mistral-7B-Instruct	1.93	1.92	2.02	2.01	2.00	1.97	2.06	2.33	2.02	2.00	1.72	2.00
Mistral-small-latest	**2.84**	**2.99**	**3.00**	2.70	**3.00**	2.80	2.62	**2.98**	2.32	**3.00**	**2.49**	**2.79**

Table 5: Average instructional alignment scores of the generated rationales. Dimension abbreviations follow those defined in Table 4. The best-performing model for each dimension is highlighted in bold.

models, Llama-3.3-70B shows the second-best performance and even achieves the lowest RMSE on the dimensions of local sufficiency and credibility. This suggests that larger models can still outperform smaller ones on specific dimensions, although the overall trend favors Mistral-small-latest.

We further examine whether providing demonstration examples improves annotation performance. The last row in Table 4 shows the performance of Mistral-small-latest under zero-shot setting. Across all evaluated dimensions, the two-shot configuration consistently yields lower RMSE than the zero-shot setting. This result indicates that including a small number of annotated examples helps the model better calibrate its scoring behavior, leading to more accurate predictions.

Overall, the results indicate that structural rationale prompting substantially improves the alignment between LLM predictions and human annotations. Across multiple models and evaluation metrics, the prompting framework consistently increases agreement and reduces prediction error, suggesting that guiding LLMs through structured reasoning can enhance their reliability as automated evaluators of argument quality.

5 Instructional Alignment

The second research question investigates instructional alignment, focusing on the quality of the rationales generated by LLMs. In particular, we examine whether the models **follow the structured evaluation instructions provided in the prompts**.

Score Difference	**Fully C.**	**Partially C.**	**Non-C.**
$\Delta = 0$	575	149	2
$\Delta = 1$	255	60	1
$\Delta = 2$	12	4	1

Table 6: Distribution of instances across instructional alignment levels (*fully compliant*, *partially compliant*, *non-compliant*) together with the corresponding prediction errors of `Mistral-small-latest`.

To this end, we randomly sample five arguments for each quality dimension and generate rationales using the four models described in Section 3.4. A total of 5 samples x 11 dimensions x 4 models = 220 annotations were conducted. Two human experts from the authors then evaluate how well each generated rationale follows the given instructions using a three-point Likert scale (1 = non-compliant, 2 = partially compliant, 3 = fully compliant). In addition to human evaluation, we prompt an independent LLM (GPT-5.2) to perform the same assessment. The two human annotators achieve a Krippendorff's α of 0.69, indicating substantial agreement. When the LLM judge is included as a third annotator, the overall inter-annotator agreement is 0.56, suggesting moderate agreement.

Table 5 reports the instructional alignment scores assigned by GPT-5.2 across all quality dimensions. Overall, the results indicate that the evaluated models generally follow the structured instructions well. Several models even achieve perfect scores in specific dimensions—for example, Mistral-small-latest obtains full compliance in local sufficiency,

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Llama-3.3-70B-Instruct	**2.87**	**2.95**	**2.99**	**3.00**	**3.00**	2.96	**2.91**	**2.99**	**2.98**	**2.98**	**2.99**	**2.96**
GPT-3.5-Turbo	2.77	2.74	2.80	2.70	2.84	2.76	2.79	2.77	2.92	2.85	2.91	2.80
Mistral-7B-Instruct	2.70	2.71	2.80	2.82	2.81	2.91	2.75	2.96	2.95	2.85	2.58	2.80
Mistral-small-latest	2.79	2.93	**2.99**	2.93	2.98	**2.99**	2.77	2.96	**2.98**	**2.98**	2.86	2.92

Table 7: Faithfulness alignment scores across models and dimensions, computed from the absolute difference between model-predicted scores and faithful scores inferred from the generated rationales. Dimension abbreviations follow those defined in Table 4. The best-performing model for each dimension is highlighted in bold.

Models	LA	LR	LS	Cr	Em	Cl	Ap	Ar	GA	GR	GS	Avg.
Mistral-small-latest end-to-end	**0.73**	0.71	0.71	0.66	0.70	0.67	0.67	0.74	0.66	0.68	**0.58**	0.68
Mistral-small-latest 2-step	0.90	**0.64**	**0.70**	**0.48**	**0.40**	**0.61**	**0.64**	**0.58**	**0.52**	**0.43**	**0.58**	**0.59**

Table 8: RMSE comparison between the original model-predicted scores and the faithful scores inferred from the generated rationales using the two-step evaluation pipeline. Dimension abbreviations follow those defined in Table 4. The best-performing setting for each dimension is highlighted in bold.

emotional appeal, and global relevance. Among the four models, Mistral-small-latest again achieves the strongest overall performance, ranking first in 7 out of the 11 evaluated dimensions.

Beyond model comparison, we further examine the relationship between instructional alignment and prediction accuracy. Table 6 analyzes the outputs of the best-performing model, Mistral-small-latest, by grouping instances according to their instructional alignment level. The results reveal a general tendency: rationales that more closely follow the provided instructions are associated with lower prediction errors. This observation suggests that adherence to the structured reasoning procedure may contribute to more reliable argument quality assessments.

Overall, the results indicate that modern LLMs can largely follow the structured evaluation instructions provided in the structural rationale prompts. The consistently high instructional alignment scores suggest that the generated rationales generally adhere to the intended evaluation criteria. Moreover, the observed association between higher instructional compliance and lower prediction error indicates that closely following the rational procedure may contribute to more reliable argument quality assessments.

6 Faithfulness Alignment

While the previous two research questions examine predicted scores and generated rationales separately, the third research question investigates faithfulness alignment, that is, **the internal consistency between the scores produced by a model and the rationales it generates**. To assess this relationship, we use the same randomly selected examples described in Section 5, and conducted another 220 annotations. Two human experts and one LLM judge (GPT-5.2) are shown only the generated rationales, without access to the original arguments or the model-predicted scores. Based solely on the rationales, they are asked to assign argument quality scores, which we refer to as *faithful scores*. The agreement between the two human annotators reaches a Krippendorff's α of 0.65, while the overall inter-annotator agreement across all three annotators is 0.53, indicating moderate agreement.

Using the annotations produced by GPT-5.2, we define faithfulness alignment based on the absolute difference between the faithful score and the original model-predicted score. A value of 3 indicates perfect alignment (no difference), 2 indicates a difference of one point, and 1 indicates a difference of two points. Table 7 reports the resulting faithfulness alignment scores across all dimensions and models. Overall, the predicted scores are highly consistent with the generated rationales. The lowest average faithfulness score is 2.80 for Mistral-7B-Instruct, while Llama-3.3-70B-Instruct achieves the highest score of 2.96, indicating very strong internal alignment.

Finally, we extend the analysis by asking GPT-5.2 to infer faithful scores for the entire dataset and comparing these scores with the human annotations, following the evaluation protocol used in Section 4. This procedure effectively implements a two-step evaluation pipeline: the model first generates a rationale and then derives a score from that rationale. Such a setup resembles multi-step reasoning approaches such as Least-to-Most prompting (Zhou et al., 2023). Table 8 reports RMSE for each dimension, compared with the best-performing configuration from Section 4 (`Mistral-small-latest` with structural rationale prompting). The results show that this two-step approach further reduces RMSE, suggesting that explicitly grounding the score in the generated rationale can lead to more accurate predictions.

Overall, the results indicate that the scores produced by LLMs are largely supported by their generated rationales, demonstrating strong faithfulness alignment. Moreover, deriving scores explicitly from rationales in a two-step evaluation pipeline can further improve prediction accuracy. These findings suggest that incorporating structured reasoning not only improves interpretability but can also enhance the reliability of LLM-based argument quality assessment.

7 Conclusion

In this paper, we proposed a three-level audit framework for evaluating the reliability of LLMs as annotators for argument quality assessment, covering surface alignment, instructional alignment, and faithfulness alignment. To operationalize this framework, we introduced structural rationale prompting, which guides models to generate structured justifications before assigning scores. Experiments on the Dagstuhl-15512 argument quality corpus show that this prompting strategy substantially improves agreement with human annotations, while the generated rationales generally follow the intended evaluation criteria and remain highly consistent with the predicted scores. Overall, our findings highlight the importance of auditing LLM evaluators beyond surface score agreement in order to better understand the transparency and reliability of LLM-based argument evaluation.

8 Limitations

This study has several limitations. First, our experiments focus on argument quality assessment using the Dagstuhl-15512 corpus, which contains a limited number (320 instances) of annotated arguments from a specific debate dataset. As a result, the findings may not fully generalize to other domains or argumentation settings. Second, we evaluate a limited set of LLMs and prompting strategies. Although the selected models represent different model families and scales, the results may vary for other models or alternative prompting approaches.

Third, part of our evaluation relies on an LLM-based judge to assess instructional and faithfulness alignment. While we validate the judge against human annotations on a subset of instances, automated evaluation may still introduce biases or inaccuracies. Finally, the structural rationale prompts used in this work are designed primarily to facilitate the proposed auditing framework rather than to represent an optimal prompting strategy. Future work could explore alternative prompt designs and examine whether similar alignment patterns hold across different reasoning structures.

9 Ethical Concerns

The increasing use of LLMs as automated annotators raises ethical considerations regarding the role of human expertise in data annotation and corpus creation. While LLM-based annotation may reduce the cost and time required to construct labeled datasets, it should not be viewed as a direct replacement for human annotators, particularly for complex tasks such as argument quality assessment that involve nuanced judgment and contextual understanding. The goal of this work is therefore not to advocate replacing human annotators, but rather to examine whether and how LLM outputs align with human evaluation criteria.

Another concern is the potential over-reliance on LLM-generated annotations in downstream research. If LLM outputs are used without careful validation, systematic biases or reasoning errors may propagate into training data and evaluation benchmarks. Our proposed auditing framework aims to mitigate this risk by encouraging researchers to evaluate LLM annotators along multiple alignment dimensions, including their reasoning processes. By promoting transparency in how LLM evaluators generate scores and explanations, we hope to support more responsible use of LLMs in annotation and evaluation pipelines.

Acknowledgments

This paper is based in part on results obtained from the AIST policy-based budget project “R&D on Generative AI Foundation Models for the Physical Domain.” This work was also supported by the AISafety Project, funded by the BMBF under grant proposal 05D23PD1, and by the State of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence.

References

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse (George A. Kennedy, Translator)*. Clarendon Aristotle series. Oxford University Press.

J Anthony Blair. 2011. *Groundwork in the theory of argumentation: Selected papers of J. Anthony Blair*, volume 21. Springer Science & Business Media.

WK Chan, YT Yu, Jacky W Keung, and Victor CS Lee. 2023. Toward ai-assisted exercise creation for first course in programming through adversarial examples of ai models. In *2023 IEEE 35th International Conference on Software Engineering Education and Training (CSEE&T)*, pages 132–136. IEEE.

Wei-Fan Chen, Milad Alshomary, Maja Stahl, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2024. Reference-guided style-consistent content transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13754–13768.

Wei-Fan Chen, Zhixue Zhao, Akbar Karimi, and Lucie Flek. 2025. Explainable hallucination through natural language inference mapping. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1888–1896.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational linguistics*, 45(4):765–818.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

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.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Conference on Advances in Robust Argumentation Machines*, pages 129–146. Springer.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. Here's my point: Joint pointer architecture for argument mining. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1364–1373.

Zahra Rahimi, Diane J Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. 2014. Automatic scoring of an analytical response-to-text assessment. In *International conference on intelligent tutoring systems*, pages 601–610. Springer.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Maja Stahl, Timon Ziegenbein, Joonsuk Park, and Henning Wachsmuth. 2025. Arginstruct: specialized instruction fine-tuning for computational argumentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11103–11127.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment-new datasets and methods. 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 5625–5635.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

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, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, and 1 others. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

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.

Yuxiao Ye and Simone Teufel. 2021. End-to-end argument mining as biaffine dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 669–678.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

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.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and 1 others. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Appendix

A.1 Full Example

Table 9 provides an argument together with its issue and stance. Using this argument, the tested four models generate the rationales and predict the quality score as shown in Table 10 and 11. Besides, the alignment scores by GPT-5.2 are also included in the tables.

Human Annotation: 1 (lowest)

Issue: ban-plastic-water-bottles
Stance: no-bad-for-the-economy
Argument: Western Europe consumes almost 50% of the world's bottled water. Many claim that this is due to European culture, since the continent has had very polluted waters due to agriculture and industry dating back to the Industrial Revolution. The water bottles they buy help us.

Table 9: An example argument together with its issue and stance.

Llama-3.3-70B-Instruct
Predicted Score: 1
Instructional Alignment: 3
Faithfulness Alignment: 3

Rationales:
1. **Conclusion**: The argument aims to support the stance "no-bad-for-the-economy" — i.e., banning plastic water bottles is not bad for the economy. However, the stated conclusion is not clearly articulated. Instead, the argument makes a claim about Western Europe's bottled water consumption and cultural reasons for it, concluding with "The water bottles they buy help us." This is vague and does not logically connect to the economic impact of a ban.
2. **Premises**:
- Western Europe consumes nearly 50% of the world's bottled water.
- This consumption is attributed to historical pollution from agriculture and industry (dating to the Industrial Revolution).
- The purchase of bottled water by Europeans "helps us" (implying economic benefit to others, likely producers or exporters).
3. **Evaluation of Logical Connections**:
- The premises do not address the economic consequences of banning plastic water bottles.
- The claim that "the water bottles they buy help us" is ambiguous: it's unclear who "we" refers to, what kind of help is meant (e.g., export revenue, job creation), and how this relates to a ban.
- Even if Western Europe's consumption supports some industries, this does not justify the claim that a ban would be *bad for the economy*. In fact, a ban might shift demand to alternatives (e.g., reusable bottles, filtered tap water), which could still support economic activity.
- There is no reasoning linking the current consumption pattern to the economic viability of maintaining plastic bottle use under a ban.
4. **Assessment of Premise Quality**:
- The premise about Western Europe consuming 50% of bottled water is plausible and supported by data (e.g., from Euromonitor or OECD reports), so it is factually acceptable.
- The historical pollution claim is broadly consistent with known environmental history of Western Europe.
- However, the causal link between pollution and bottled water use is oversimplified and lacks nuance (e.g., many European countries have high-quality tap water but still consume bottled water due to convenience or marketing).
- The final claim — "The water bottles they buy help us" — is unsupported, vague, and lacks evidence or explanation.
5. **Global Judgment**:
- The premises fail to provide sufficient support for the conclusion that banning plastic water bottles is bad for the economy.
- There is a significant gap between the evidence presented and the economic claim being defended.
- The argument appears to confuse current consumption patterns with long-term economic impacts of regulation.
- It does not address alternatives, market shifts, or potential economic benefits of a ban (e.g., reduced waste management costs, growth in sustainable industries).
Therefore, the argument suffers from a fundamental lack of logical sufficiency. The premises are either irrelevant or too weak to justify the conclusion, despite some plausible factual claims.

GPT-3.5-Turbo
Predicted Score: 2
Instructional Alignment: 2
Faithfulness Alignment: 3

Rationales:
The premises provided in the argument are relevant and logical, linking European culture and historical context to the consumption of bottled water. However, the argument lacks depth and evidence to fully support the conclusion that water bottles help the economy. The connection between cultural history and economic benefit is not explicitly detailed, leading to a partially sufficient local sufficiency score.

Table 10: The output rationales together with the predicted local sufficiency quality score and the alignment scores by Llama-3.3-70B-Instruct and GPT-3.5-Turbo.

Mistral-7B-Instruct
Predicted Score: 2
Instructional Alignment: 2
Faithfulness Alignment: 3

Rationales:
While the premises are relevant and logically connected, they lack depth and specificity in supporting the conclusion that the Western European consumption of bottled water is not bad for the economy. The argument relies on generalizations about European culture and history without providing concrete evidence of the economic benefits of bottled water consumption. Additionally, the premises could be strengthened by including specific data or examples that demonstrate the positive impact of bottled water on the economy.

Mistral-small-latest
Predicted Score: 1
Instructional Alignment: 3
Faithfulness Alignment: 3

Rationales:
1. **Conclusion**: The argument aims to support the stance that banning plastic water bottles is bad for the economy by claiming that plastic water bottles are beneficial to the economy (implied by “The water bottles they buy help us”).
2. **Premises**:
- (1) Western Europe consumes almost 50% of the world’s bottled water.
- (2) This consumption is due to European culture, as the continent has had polluted waters due to agriculture and industry since the Industrial Revolution.
- (3) The water bottles they buy help us (presumably the global economy or the author’s country).
3. **Evaluation**:
- The premises are somewhat relevant to the conclusion but contain significant gaps in reasoning.
- Premise (1) provides a factual claim about consumption but does not directly explain how this supports the economy.
- Premise (2) attempts to justify the consumption by historical context but does not logically connect to economic benefits.
- Premise (3) is vague ("help us") and lacks specificity about how the economy benefits. It assumes a positive economic impact without evidence.
- The argument does not explain why banning plastic water bottles would harm the economy, nor does it provide data or reasoning to support the claim that bottled water is economically beneficial.
4. **Overall Judgment**:
- The premises are weak, lack depth, and fail to provide sufficient logical or evidentiary support for the conclusion. The argument relies on unsupported assumptions and broad generalizations.

Table 11: The output rationales together with the predicted local sufficiency quality score and the alignment scores by Mistral-7B-Instruct and Mistral-small-latest.

Topic-Guided Prompting for Argument Stance Classification

Bogdan Grecu and Oana Cocarascu
King's College London
{bogdan.grecu, oana.cocarascu}@kcl.ac.uk

Abstract

Stance classification is a core task in argument mining and subjectivity analysis, crucial for understanding public discourse and opinion dynamics on social media. Despite their impressive few-shot capabilities, Large Language Models (LLMs) remain sensitive to prompt construction, including the selection and ordering of in-context examples. In this paper, we propose a Topic-Guided prompting method for argument stance classification that dynamically integrates topic-specific information into the few-shot context. We evaluate our method on five LLMs across three datasets spanning formal debates and user-generated online comments. Our extensive evaluation shows that our proposed Topic-Guided prompting outperforms standard few-shot prompting and state-of-the-art example selection strategies. Further analysis indicates that our method reduces the bias towards the 'support' class observed in several models, resulting in more balanced predictions across stances and thus a more robust approach to stance classification.

1 Introduction

Argument mining is a sub-field of natural language processing (NLP) that involves automatically identifying and analysing argumentative structures from text (see Lippi and Torroni (2016); Lawrence and Reed (2019); Lauscher et al. (2022) for surveys). It typically encompasses three main tasks: 1) argument identification, which involves determining text spans that constitute arguments with respect to a given topic; 2) component classification which aims to identify claims and premises; and 3) stance classification which determines the position an argument takes towards a topic or another argument.

Stance classification, i.e. identifying whether a text expresses a viewpoint for or against a given target, has numerous applications ranging from identifying misinformation (Hardalov et al., 2022) to public health policymaking (Glandt et al., 2021) and news recommendations (Reuver et al., 2021a). As social media platforms and online forums become the primary venues for public debate, automatically identifying the stance expressed towards a given claim becomes essential for modelling argumentative interactions and understanding the dynamics of online opinion formation.

However, the wide range of topics on social media poses a challenge for robust stance classification (Allaway and McKeown, 2020). Prior works on cross-topic stance detection found that pre-trained language models often fail to generalize across diverse subjects (Cocarascu et al., 2020; Reuver et al., 2021b; Ruiz-Dolz and Iranzo-Sánchez, 2024). More recently, large language models (LLMs) have shown impressive in-context learning capabilities, enabling strong performance from a small set of input-output examples (Brown et al., 2020), and are increasingly being applied to argument mining tasks, including argument stance classification (Chen et al., 2024; Gorur et al., 2024).

Despite these efforts, one aspect that has remained largely underexplored in argument mining research is the use of topic knowledge. In contrast, incorporating topic information has proven beneficial in several natural language understanding tasks such as named entity recognition (Jansson and Liu, 2017), sentiment analysis (Seilsepour et al., 2023), summarisation (Han et al., 2024), and machine translation (Aycock and Bawden, 2024). Prior work has shown that priming strategies that preferentially select semantically or stance-similar few-shot examples can substantially influence classification outcomes in few-shot settings (Ajjour and Wachsmuth, 2025). Furthermore, Lauscher et al. (2022) emphasize the importance of integrating world and topic knowledge in argument mining tasks to enable a more nuanced interpretation of argumentative discourse.

To this end, we propose an effective approach to argument stance detection by incorporating topic

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 39–51
July 3, 2026

knowledge into few-shot prompting. Unlike prior works on stance classification which typically rely on static prompts, our method dynamically adapts the prompt at inference time based on the topic[1] most representative of the instance tested. This enables the model to exploit topic-specific cues that are crucial for accurately interpreting subjective and context-dependent arguments.

We evaluate our Topic-Guided prompting approach on five open-weights and proprietary models that vary in size: Llama-3.2-3B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Gemma-3-12B (Mesnard et al., 2024), Qwen2-7B (Bai et al., 2023), and GPT-3.5 Turbo (OpenAI). Our experiments on three of the most widely used datasets in argument stance classification, IBM-Debater (Bar-Haim et al., 2017a), UKP (Stab et al., 2018), and Perspectrum (Chen et al., 2019), show that our method outperforms standard few-shot prompting as well as state-of-the-art example selection strategies for few-shot learning such as kNN (Liu et al., 2022), TopK+ConE (Peng et al., 2024), and perplexity estimation (Gonen et al., 2023). Moreover, our Topic-Guided prompting approach helps mitigate the tendency of several models to exhibit bias towards one label, resulting in a more robust approach to stance classification

To summarize, our **contributions** are:

- We propose Topic-Guided prompting, a dynamic few-shot prompting strategy for argument stance classification that actively conditions the model on topic-specific information at inference time, enhancing the interpretation of context-dependent arguments.

- We conduct a comprehensive evaluation using five LLMs on three datasets, showing that our proposed method consistently outperforms standard few-shot prompting and advanced example selection techniques. Furthermore, we show that our method is computationally efficient, requiring fewer LLM queries compared to baselines.

- We demonstrate that our approach mitigates model bias towards specific labels (particularly the tendency to over-predict the 'support' class), thereby reducing the performance gap between opposing stances compared to baseline methods.

- We provide a cross-dataset evaluation to assess our method's generalisability across different domains.

[1]We refer to 'topic' as the explicit central claim or controversial issue under discussion, as defined in standard argument mining (Stab et al., 2018), rather than latent topics derived via unsupervised topic modelling.

2 Related Work

2.1 Argument Mining

Prior to the advent of LLMs, research in argument mining relied on feature engineering to deploy statistical machine learning algorithms (Lippi and Torroni, 2016; Lawrence and Reed, 2019). Similarly, early works on argument stance detection employed features commonly used in general NLP and sentiment analysis tasks (Hasan and Ng, 2014; Rajendran et al., 2018; Sobhani et al., 2017), specific argumentative features (Somasundaran and Wiebe, 2010), or contextual features (Bar-Haim et al., 2017b). More recently, studies on stance detection deployed deep learning models such as hierarchical attention networks (Sun et al., 2018), fine-tuned language models such as BERT (Durmus et al., 2019), or multi-lingual BERT and transfer learning (Toledo-Ronen et al., 2020).

Given the remarkable success of LLMs in understanding and generating human-like language, recent works have started evaluating their capabilities in several argument mining tasks such as argument component identification (Kashefi et al., 2023), classifying relations between argument components (Otiefy and Alhamzeh, 2024), stance classification (Gorur et al., 2024), as well as argument generation (Chen et al., 2024). Whilst some works focus on fine-tuning LLMs for various tasks such as classifying argument components and argument relations (Cabessa et al., 2025; Yuan et al., 2024), other works examine LLMs in zero-shot and few-shot prompting settings (Chen et al., 2024; Gorur et al., 2025; Bezou-Vrakatseli et al., 2025).

2.2 Topic-specific Methods in NLP

Several works have explored the use of topic-specific methods in various NLP tasks to capture these nuances. For instance, Ding et al. (2022) analyzed the importance of topic-specific information for argument mining tasks. The authors used a topic-based modelling approach to cluster the data, and then trained argument mining models with different combinations of topic-specific and topic-independent data. Allaway and McKeown

(2020) developed a model for zero-shot stance detection that exploits information about topic similarity through generalized topic representations, addressing the challenge of shifting contexts in online debates. Xiao et al. (2023) used topic-based prompts for deep passage retrieval. Here, the prompts used are pieces of text that hold task-specific information. Each document is assigned a unique prompt that aligns with its semantic and topical diversity. These prompts are then used by a pre-trained language model to obtain representations that align with their documents' topic distributions.

Similar to prior work, we recognise the importance of incorporating topic-specific information for capturing the nuances of subjectivity in NLP tasks. Our topic-guided prompting method directly integrates topic knowledge into few-shot prompting for argument stance classification.

2.3 Few-shot Learning Example Selection Methods

Exploring how to better create prompts for LLMs and how to better select examples for few-shot learning has been the focus of several works. Liu et al. (2022) proposed a method for selecting few-shot learning examples for several natural language understanding and generation tasks. In particular, a subset of examples is sampled randomly from the dataset and then used as a selection pool for the examples in the few-shot learning prompt. The examples from the training set sent to the LLM as context for the evaluation of the test example are selected as the K nearest neighbours, in the embedding space, of the test example. Peng et al. (2024) expanded on this by increasing the number of examples returned by the K nearest neighbours algorithm and then used the examples that would yield the smallest conditional entropy together with the test example. Thus, their method, TopK + ConE, enhances the models' understanding of the test instance by reducing the conditional entropy of the test instance under the inference model.

Gonen et al. (2023) used perplexity estimation to select the prompt. While their method is used for zero-shot learning, it can be adapted to select examples for few-shot learning. Rubin et al. (2022) used unsupervised learning and contrastive learning in order to train a dense retriever that is then used, at inference, to select the best performing examples from the training set for a particular test instance. Zhang et al. (2022) proposed a reinforcement learning framework based on Q-learning to select the few-shot learning examples.

Within the argument mining domain, Ajjour and Wachsmuth (2025) investigated priming strategies for few-shot stance classification. The authors compared selecting examples based on semantic similarity versus stance similarity, introducing affective priming (leveraging stance-similar instances), semantic priming (utilizing semantically similar instances) and distinct-k (ensuring diverse instances via clustering). Their findings highlight the importance of tailoring selection strategies to the argumentative nature of the task, revealing that stance similarity can significantly steer model predictions.

The proposed Topic-Guided method distinguishes itself from example selection strategies such as kNN, TopK+ConE, and Perplexity, by explicitly incorporating the topic information provided with the data. While these methods rely primarily on semantic similarity or probability metrics to select examples for each individual instance, Topic-Guided prompting uses this given topic to condition the model on the known topic. This allows the model to effectively exploit topic-specific cues that are crucial for interpreting context-dependent arguments. This distinction is also evident when comparing our approach to recent priming strategies explored by Ajjour and Wachsmuth (2025). While their distinct-k method aims to diversify the prompt by clustering instances based on semantic similarity, and their affective priming strategy retrieves few-shot examples with a similar stance to the target instance, neither approach incorporates topic information, highlighting a key advantage of our Topic-Guided prompting.

3 Topic-Guided Few-shot Prompting

In this section, we introduce our dynamic few-shot prompting strategy for argument stance classification, which conditions the model on topic-specific information at inference time. In standard few-shot prompting, the model is provided with a small set of labelled examples to predict the class during inference, leveraging in-context learning (Dong et al., 2024). However, most prior work focuses on defining task instruction, with the examples provided in the task being randomly sampled (Chen et al., 2024; Gorur et al., 2025).

In the following, we describe our new approach for prompting LLMs for stance classification, Topic-Guided prompting, which adapts automatically during few-shot learning based on the

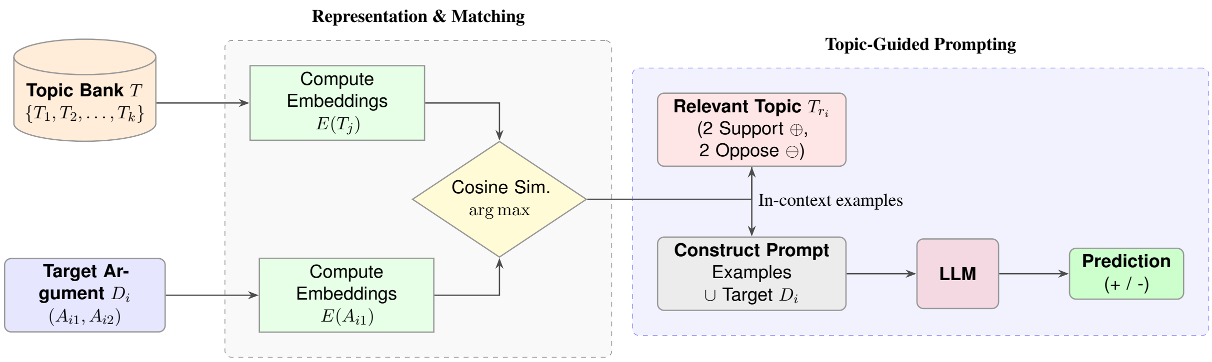


Figure 1: A procedural overview of our Topic-Guided prompting approach. Target arguments and topic-bank entries are embedded and matched via cosine similarity to retrieve relevant in-context examples for the LLM.

topic of the argument given at inference time, thus enabling more informed and context-sensitive predictions.

Consider a dataset $D = \{D_1, D_2 \dots D_n\}$ consisting of n instances. Each instance $D_i = (A_{i1}, A_{i2})$ is a pair comprising a central claim A_{i1} and a target argument A_{i2}. The aim of the argument stance classification task is to determine whether argument A_{i2} *supports* or *opposes* A_{i1}. Given a set of topics forming a topic-bank $T = \{T_1, T_2 \dots T_k\}$, where each T_j represents a unique motion or central claim extracted from a source dataset, we first create a set of prompts[2] for each topic containing two examples for each class, *support* and *oppose*, relevant to the topic. Our goal is to find the most relevant topic T_r for each D_i.

For each $T_i \in T$, we compute the sentence embeddings[3] using a self-supervised contrastive learning objective (Wang et al., 2020). Similarly, we compute the embeddings E for each $A_{i1} \in D_i$ whose class a model needs to infer. Note that A_{i1} and A_{i2} are related to the same topic as we are considering a binary classification task. Then, we select the most relevant topic $T_{r_i} \in T$ for each D_i using cosine similarity $T_{r_i} = \arg\max_j(cos\_sim(E(A_i), E(T_j)))$.

Figure 1 shows how examples are selected in our Topic-Guided prompting method. This retrieval step enhances the method's applicability in real-world settings. While an exact topic match may not be available for every unseen argument, our system dynamically maps each input to the most semantically similar topic within its topic repository T. This enables the model to base its reasoning on the most relevant available context, rather than relying on generic examples.

[2]Prompts can be found in Appendix A.

[3]To calculate the sentence embeddings, we used the pre-trained model all-MiniLM-L6-v2 from sentence-transformers.

4 Experiments

In this section, we give an overview of the datasets and large language models used in our evaluation, as well as the baselines used to compare our Topic-Guided prompting method.

4.1 Datasets

We evaluate our method on three commonly used datasets in argument stance classification: IBM-Debater (Bar-Haim et al., 2017a), UKP (Stab et al., 2018), and Perspectrum (Chen et al., 2019).

IBM-Debater The dataset is sourced from the International Debate Education Association and focuses on structured, high-quality argumentation regarding controversial motions. Unlike unstructured text, these entries represent formal claims that explicitly support or contest a central topic, such as the regulation of media or public policy.

UKP In contrast to formal debates, the UKP dataset consists of arguments extracted from user-generated online comments, capturing the raw and often informal nature of public discourse. It centres on eight highly polarized societal issues such as gun control and the death penalty, allowing for an in-depth analysis of sentiment within specific domains. As one of the arguments in the UKP dataset is represented by the topic, in our experiments we create a position claim from the topic, specifically *topic is good*. For example, "Reproductive cloning would diminish the sense of uniqueness of an individual" opposes the argument "Cloning is good".

Perspectrum This dataset consists of crowd-sourced claims from the community debate portal *debate.org*. It covers a wide variety of subjects,

Dataset	# Topics	# Examples	# Support examples	# Oppose examples
IBM	55	2394	1325	1069
UKP	8	11032	4897	6135
Perspectrum	18	10231	5279	4952

Table 1: Number of topics and distribution of support/oppose arguments in the three datasets.

Dataset	Topic	Central Claim (A_{i1})	Target Argument (A_{i2})	Relation
IBM	the sale of violent video games to minors	This house believes that the sale of violent video games to minors should be banned	Violent video games increase the violent tendencies among youth	support
	the sale of violent video games to minors	This house believes that the sale of violent video games to minors should be banned	No conclusive link was found between video game usage and violent activity	oppose
UKP	cloning	Cloning is good	reproductive cloning could aid in genetic research.	support
	cloning	Cloning is good	reproductive cloning would diminish the sense of uniqueness of an individual.	oppose
Perspectrum	politics	We should restrict violent video games	Violent video games are detrimental to social interaction.	support
	politics	We should restrict violent video games	Gun violence is less prevalent in countries with high video game use.	oppose

Table 2: Examples from the three datasets: IBM-Debater, UKP, and Perspectrum.

with every claim explicitly labelled with its stance support or oppose relative to the parent topic. For our experiments we have excluded the data items that had the topic "Unknown".

Table 1 shows statistics of the datasets we used while Table 2 shows examples from the datasets.

In the benchmark datasets utilised, each instance is generally linked to a single main controversial topic. In cases where topic labels were missing, ambiguous, or noisy (such as the 'Unknown' category in the Perspectrum dataset), the instances were excluded during preprocessing to ensure a rigorous evaluation of our Topic-Guided approach. However, to examine how our method handles unseen or out-of-domain topics that may arise in real-world applications, we conduct a dedicated cross-dataset evaluation, detailed in Section 5.2.

4.2 Large Language Models

We evaluate five open-weights and proprietary models that vary in size: Llama-3.2-3B (Grattafiori et al., 2024), Mistral-7B (Jiang et al., 2023), Gemma-3-12B (Mesnard et al., 2024), Qwen2-7B (Bai et al., 2023), and GPT-3.5 Turbo (OpenAI) and compare the performance of our Topic-Guided prompting with standard few-shot prompting. We conducted inference on the complete datasets reported in Table 2, rather than on sampled subsets. In all experiments, we employed Greedy Decoding to ensure deterministic and reproducible outputs. Consequently, we report the results of a single run per instance, as the deterministic nature of the decoding eliminates generation variance. Experiments were run on two NVIDIA A100 GPUs.

4.3 Baselines

We compare our Topic-Guided prompting method with standard few-shot prompting and five state-of-the-art methods for selecting examples for in-context learning: k-Nearest Neighbours (kNN) (Liu et al., 2022), TopK + Conditional Entropy (ConE) (Peng et al., 2024), Perplexity estimation (Gonen et al., 2023), distinct-k and affective priming (Ajjour and Wachsmuth, 2025) (see Section 2.3 for an overview).

For Topic-Guided few-shot prompting, we select the first two pairs of supporting arguments and the first two pairs of opposing arguments for each topic, following their original order in the dataset. This selection ensures strict reproducibility and eliminates performance variance that comes from random sampling. After the topic is determined, the corresponding prompt is fixed, and no further dynamic re-selection of examples is performed during evaluation. For kNN, ConE, and perplexity estimation, we set the sample size equal to the number of arguments selected for the topic-specific prompts constructed in our approach. As such, for both supporting arguments and opposing arguments, the initial sample size is $2 * N_T$, where N_T denotes the total number of topics in the dataset.

To ensure a rigorous and fair evaluation, we excluded all argument pairs used as in-context examples for a given topic from the test set for that topic, thereby preventing any overlap between the demonstration and evaluation data.

Standard few-shot prompting Standard few-shot prompting follows the conventional in-context learning paradigm, where demonstration examples are randomly sampled. In our experiments, we sample a fixed set of examples (i.e. two support, two oppose) at the start of the experiment. This single set of examples is then used as a static context for all test instances, irrespective of the topic of the argument being classified.

Selecting examples using kNN We use the approach of Liu et al. (2022) to select examples using kNN. In particular, 2*N_T supporting and 2*N_T opposing arguments are randomly selected. For each argument to be evaluated, we select the two nearest neighbours, in the embedding space, from the supporting arguments, and the two nearest neighbours, in the embedding space, from the opposing arguments. The few-shot learning prompt then consists of these two supporting and two opposing arguments for the example we are evaluating.

Selecting examples using TopK+ConE Following Peng et al. (2024), we retrieve the four nearest neighbours in the embedding space from the supporting arguments, and the four nearest neighbours in the embedding space from the opposing arguments. Then, we calculate the conditional entropy between each supporting argument and the argument we are evaluating, and select the two arguments with the lowest conditional entropy. We apply the same procedure to the opposing arguments, selecting the two instances that will be used in the few-shot prompt for the arguments we are evaluating. We exclude the closed-source model (GPT) from this experiment, as it does not provide access to token probabilities required to compute conditional entropy.

Selecting examples using perplexity estimation We adopt perplexity estimation following Gonen et al. (2023) to select the examples used in the prompt. First, we randomly sample 2*N_T supporting and 2*N_T opposing arguments, where N_T denotes the number of topics in the dataset. We chose this sampling size to ensure a similar number of examples to that used in the Topic-Guided prompting. Then, we select the two supporting arguments that yield the lowest perplexity when considered with the argument being evaluated. We apply the same procedure to the opposing arguments, selecting the two instances with the lowest perplexity. Finally, the resulting four arguments are then used in the few-shot prompt. We exclude the closed-source model (GPT) from this evaluation, as the model does not expose the token probabilities required to calculate the perplexity.

Selecting examples using distinct-k Following Ajjour and Wachsmuth (2025), we group the training instances into k clusters based on semantic similarity using agglomerative clustering. The most central representative of each cluster is then used as a prime to ensure a diverse set of few-shot examples that cover the semantic space. We set $k = 4$ to ensure that the prompt contains the same number of examples as in the Topic-Guided method.

Selecting examples using affective priming As in Ajjour and Wachsmuth (2025), we use contrastive learning embeddings to capture stance similarity. We select the k training instances that are most stance-similar to the target instance to serve as primes. For this experiment, we first split the datasets into training and test sets. The training set is used to train the contrastive embedding model that captures stance similarity and also serves as the candidate pool for example selection. For each target argument in the test set, we compute the cosine similarity between its contrastive embedding and those of all arguments in the training set. During inference, for each new example we retrieve the top k instances with the highest stance similarity and use them to construct the few-shot prompt. To keep the prompt size consistent with our Topic-Guided method we set $k = 4$.

5 Results and Discussion

We evaluate the effectiveness of our Topic-Guided prompting method against standard few-shot prompting and state-of-the-art example selection strategies. We report results using the weighted average F_1 score. Furthermore, we also examine the generalisability of our method through cross-dataset experiments.

5.1 How does Topic-Guided Prompting compare to other methods?

Table 3 shows the F_1 scores for the two classes, *support* and *oppose*, as well as the average F_1 score.

	LLM	Topic-Guided	Standard	kNN	TopK+ConE	Perplexity est.	Distinct-k	Affective priming
IBM	Llama3	69.8 / 64.4 / 67.3	69.6 / 54.8 / 63.7	72.3 / 58.4 / 66.7	72.5 / 58.4 / 66.9	67.5 / 61.3 / 64.7	68.8 / 64.9 / 67.1	67.6 / 54.8 / 60.9
	Gemma	**94.2** / **92.7** / **93.5**	93.4 / 92.2 / 92.8	92.2 / 91.0 / 91.7	92.1 / 90.9 / 91.6	93.6 / 92.6 / 93.1	93.1 / 92.3 / 92.8	92.4 / **92.7** / 92.5
	Qwen	78.2 / 80.3 / 79.1	77.9 / 80.0 / 78.8	76.5 / 79.1 / 77.6	75.8 / 78.9 / 77.1	76.5 / 79.6 / 78.5	81.9 / 82.2 / 82.0	79.8 / 83.7 / 81.6
	Mistral	82.9 / 82.5 / 82.7	79.2 / 78.5 / 78.9	81.3 / 79.6 / 80.5	81.6 / 80.0 / 80.8	79.5 / 80.0 / 79.7	80.9 / 81.7 / 81.2	83.7 / 85.0 / 84.3
	GPT-3.5	91.4 / 89.6 / 90.6	90.1 / 87.5 / 89.0	79.1 / 78.7 / 78.9	-	-	90.3 / 89.1 / 89.8	83.5 / 85.1 / 84.3
UKP	Llama3	66.4 / 66.8 / 66.6	63.0 / 59.8 / 61.5	66.9 / 65.3 / 66.1	65.6 / 62.9 / 64.3	65.6 / 70.3 / 68.1	64.2 / 67.3 / 65.9	63.1 / 49.3 / 57.0
	Gemma	83.9 / **88.2** / **86.3**	74.6 / 74.1 / 74.3	81.3 / 85.2 / 83.5	80.9 / 85.1 / 83.2	79.3 / 84.2 / 82.0	**87.0** / 84.2 / 85.6	82.1 / 85.3 / 83.9
	Qwen	73.1 / 83.9 / 79.1	71.9 / 82.0 / 77.5	74.1 / 83.3 / 79.2	73.5 / 82.8 / 78.7	71.0 / 81.7 / 77.0	74.6 / 84.0 / 79.8	74.1 / 81.5 / 78.4
	Mistral	76.9 / 83.7 / 80.9	75.0 / 82.8 / 79.6	75.3 / 82.7 / 79.7	74.9 / 83.0 / 79.7	73.4 / 83.2 / 79.4	75.1 / 85.3 / 80.8	75.9 / 81.7 / 79.1
	GPT-3.5	73.6 / 83.2 / 79.0	73.3 / 82.3 / 78.3	71.6 / 80.9 / 76.2	-	-	71.8 / 83.6 / 78.4	76.1 / 80.5 / 78.3
Perspectrum	Llama3	72.3 / 58.7 / 65.7	58.1 / 69.9 / 63.8	72.7 / 60.5 / 66.7	68.5 / 56.9 / 62.9	70.3 / 46.0 / 58.5	70.9 / 71.0 / 70.9	75.1 / 68.1 / 71.7
	Gemma	**93.9** / **93.7** / **93.8**	91.7 / 92.0 / 91.8	93.8 / 93.5 / 93.6	93.0 / 92.8 / 92.9	93.2 / 92.8 / 93.0	93.5 / 93.4 / 93.4	93.3 / 93.1 / 93.2
	Qwen	84.9 / 88.0 / 86.5	80.8 / 85.2 / 83.0	84.1 / 86.2 / 85.1	83.4 / 85.9 / 84.6	84.5 / 85.8 / 85.1	80.3 / 89.5 / 86.0	86.0 / 86.2 / 86.1
	Mistral	90.6 / 90.7 / 90.7	85.0 / 86.9 / 86.1	90.4 / 90.6 / 90.5	90.2 / 90.5 / 90.4	90.8 / 90.5 / 90.6	88.0 / 88.6 / 88.3	89.6 / 89.7 / 89.6
	GPT-3.5	81.8 / 83.2 / 82.4	68.4 / 78.6 / 73.3	81.5 / 83.0 / 82.2	-	-	81.0 / 80.6 / 80.8	84.3 / 83.9 / 84.1

Table 3: F_1 scores for support / oppose / both classes. In bold, the best F_1 scores for each dataset.

Topic-Guided prompting vs standard few-shot prompting Our proposed approach consistently outperforms standard few-shot prompting in terms of average F_1. Moreover, our method proves to be particularly effective for models that are biased towards one of the classes. Specifically, for models where the F_1 score for the *oppose* class is lower than for the *support* class, our Topic-Guided prompting method significantly reduces the difference between these scores.

Topic-Guided prompting vs kNN Our method outperforms the kNN selection strategy in terms of average F_1 in 13 out of the 15 experimental settings. Furthermore, our method is better at mitigating model bias towards specific labels compared to the kNN selection strategy. In particular, the average difference between the F_1 score for the support and oppose classes is 4.1% for the Topic-Guided setup, compared to 4.5% for the kNN setup, indicating improved class balance.

Topic-Guided prompting vs TopK+ConE Table 3 indicates comparable performance between the kNN and the TopK+ConE selection strategies. Although TopK+ConE shows a slight improvement in average F_1 over kNN, Topic-Guided prompting consistently outperforms this selection strategy across all 12 experiments. Furthermore, our Topic-Guided prompting is better at dealing with models that are biased towards one of the labels compared to TopK+ConE. The average difference between the F_1 score for the support and oppose classes for the TopK+ConE selection strategy is 4.9% compared to 4.1% for Topic-Guided prompting.

We also note that inference for TopK+ConE is slower than for Topic-Guided prompting. This can be attributed to the number of queries made to the LLM. While for Topic-Guided prompting, a single query to the LLM is required in order to obtain the label of the argument we are evaluating, for TopK+ConE, an additional of eight queries (one for each of the four supporting and four opposing arguments returned by TopK) are made in order to compute the conditional entropy of arguments with respect to the argument that is evaluated.

Topic-Guided prompting vs perplexity estimation Topic-Guided prompting outperforms perplexity estimation in 10 of the 12 experiments run. Moreover, the Topic-Guided prompting method demonstrates greater robustness to label bias compared to perplexity estimation. The average difference between the F_1 score for the support and oppose classes for the perplexity estimation selection strategy is 5.6% compared to 4.1% for Topic-Guided prompting. This method represents the slowest baseline as, for a single argument, in order to select the examples used in the few-shot prompt, one LLM query is required for every candidate argument in the selection pool.

Topic-Guided prompting vs distinct-k Our method generally achieves stronger performance across most experiments. For instance, on the IBM dataset, Topic-Guided prompting outperforms distinct-K for Gemma (93.5% vs 92.8%), Mistral (82.7% vs 81.2%), and GPT-3.5 (90.6% vs 89.8%). Similarly, on the UKP dataset, our approach yields better average F_1 scores for almost all models, including Gemma (86.3% vs 85.6%) and Llama3 (66.6% vs 65.9%). Although distinct-k occasionally performs competitively in certain configurations, it does not demonstrate the same level of consistent performance as our Topic-Guided approach across models and datasets.

	Prompting	Llama3	Mistral	Qwen	GPT-3.5	Gemma
IBM	Strongest Baseline	Distinct-k 68.8 / 64.9 / 67.1	Affective 83.7 / 85.0 / 84.3	Distinct-k 81.9 / 82.2 / 82.0	Distinct-k 90.3 / 89.1 / 89.8	Perplexity 93.6 / 92.6 / 93.1
	Topic-Guided (IBM)	69.8 / 64.4 / 67.3	82.9 / 82.5 / 82.7	78.2 / 80.3 / 79.1	91.4 / 89.6 / 90.6	94.2 / 92.7 / 93.5
	Topic-Guided (UKP)	71.5 / 43.2 / 59.1	79.4 / 77.8 / 78.7	72.7 / 76.2 / 74.2	84.2 / 82.7 / 83.5	87.1 / 85.3 / 86.3
	Topic-Guided (Persp)	71.0 / 46.7 / 60.3	78.6 / 79.9 / 79.2	77.4 / 77.7 / 77.5	86.4 / 85.2 / 85.9	86.9 / 85.4 / 86.2
UKP	Strongest Baseline	Perplexity 65.6 / 70.3 / 68.1	Distinct-k 75.1 / 85.3 / 80.8	Distinct-k 74.6 / 84.0 / 79.8	Distinct-k 71.8 / 83.6 / 78.4	Distinct-k 87.0 / 84.2 / 85.6
	Topic-Guided (UKP)	66.4 / 66.8 / 66.6	76.9 / 83.7 / 80.9	73.1 / 83.9 / 79.1	73.6 / 83.2 / 79.0	83.9 / 88.2 / 86.3
	Topic-Guided (IBM)	66.5 / 70.4 / 68.7	79.1 / 82.2 / 80.8	73.5 / 83.4 / 79.0	81.7 / 87.3 / 84.9	82.8 / 87.8 / 85.6
	Topic-Guided (Persp)	61.6 / 59.5 / 60.4	79.5 / 86.5 / 83.4	70.9 / 83.4 / 77.8	78.3 / 86.1 / 82.7	81.9 / 87.9 / 85.2
Perspectrum	Strongest Baseline	Affective 75.1 / 68.1 / 71.7	Perplexity 90.8 / 90.5 / 90.6	Affective 86.0 / 86.2 / 86.1	Affective 84.3 / 83.9 / 84.1	kNN 93.8 / 93.5 / 93.6
	Topic-Guided (Persp)	72.3 / 58.7 / 65.7	90.6 / 90.7 / 90.7	84.9 / 88.0 / 86.5	81.8 / 83.2 / 82.4	93.9 / 93.7 / 93.8
	Topic-Guided (IBM)	73.7 / 64.9 / 69.4	90.4 / 88.2 / 89.4	82.9 / 85.6 / 84.2	87.5 / 87.2 / 87.4	94.3 / 93.9 / 94.1
	Topic-Guided (UKP)	69.3 / 40.7 / 55.4	80.1 / 83.6 / 81.7	68.4 / 78.6 / 73.3	86.9 / 86.5 / 86.7	93.4 / 92.8 / 93.1

Table 4: F_1 scores for support / oppose / both classes for the cross-dataset experiments. The "Strongest Baseline" rows indicate the best performing baseline (amongst Standard, KNN, TopK+ConE, Perplexity, Distinct-k, or Affective Priming) for each model-dataset combination in Table 3.

Topic-Guided prompting vs affective priming The results indicate that Topic-Guided prompting generally outperforms affective priming across most models on the UKP and IBM datasets (e.g., Gemma achieves 93.5% vs 92.5% on IBM). However, affective priming is highly effective on the Perspectrum dataset, setting the strongest baseline for Llama3 (71.7%), and GPT-3.5 (84.1%). Despite these strong results on Perspectrum, Topic-Guided prompting exhibits greater consistency and achieves higher average performance across datasets. We also note that the inference speed for affective priming is lower compared to our Topic-Guided approach because retrieving the top k instances requires calculating the cosine similarity between the contrastive embedding of the new argument and all arguments in the training split.

5.2 Cross-dataset Evaluation

We test the generalisability of our proposed method and its ability to handle unseen topics by conducting a cross-dataset evaluation. Specifically, we prompt the LLMs using examples retrieved based on topics from one dataset (e.g. UKP) while evaluating arguments from a different dataset (e.g. IBM). This setup evaluates the method's robustness when the available topic repository does not fully align with the domain of the test instances.

Table 4 shows that the effectiveness of the Topic-Guided prompting is sensitive to the alignment between the source topics and the evaluation data. Testing on the IBM dataset sees a decrease in performance when using topics from UKP, with scores falling below the strongest baselines. This is likely attributed to the limited number of topics in UKP (8 topics), which leads to the retrieval of less semantically relevant examples for the broader range of topics in the IBM dataset (55 topics).

However, the method remains robust in other cross-dataset scenarios, and, in some cases, even outperforms the strongest baselines despite the domain shift. Notably, when evaluating on the Perspectrum dataset, using topics from IBM often leads to better performance compared to the best baselines. For example, GPT-3.5 achieves an average F_1 of 87.4% using IBM topics, significantly exceeding its strongest baseline, affective priming (84.1%). These results suggest that the diversity of topics in the IBM dataset provides a robust set of examples that can generalize well to other domains.

To assess the difficulty of this transfer task, we analyzed the topic overlap between datasets. The datasets had different numbers of topics and while some broad themes were shared (e.g. education or politics), for those, the specific motions and central claims (A_{i1}) were distinct across the three datasets.

6 Conclusion

In this paper, we introduced a novel approach to argument stance classification, namely Topic-Guided prompting, designed to better capture the context-dependent nature of argumentative discourse. Our approach actively conditions the model on topic-specific information at inference time, enhancing the interpretation of context-dependent arguments.

We evaluated our method on five LLMs and three widely used datasets: IBM-Debater (structured argumentation), UKP (user-generated comments from online forums), and Perspectrum (crowd-sourced claims from a debate portal). The results show that our approach consistently outperforms standard few-shot prompting methods and state-of-the-art example selection strategies while being equally or less computationally intensive. Importantly, our analysis revealed that Topic-Guided prompting effectively reduces the bias towards the 'support' class observed in some models.

Our results offer a new perspective for general-purpose argument mining. Rather than relying on models that treat all inputs uniformly without considering context, systems would benefit from maintaining a topic repository, i.e. a diverse collection of known issues. When encountering a novel argument, aligning it to the closest topic in this repository provides more reliable stance cues than randomly selecting examples for prompting.

There are several avenues for future work. First, we plan to evaluate our method on other datasets and other argument mining tasks, such as component identification. In addition, we plan to investigate the impact of varying the number of in-context examples as well as pooling multiple datasets to construct a heterogeneous, cross-domain topic bank. Finally, we plan to expand our study to multi-lingual data.

Limitations

This work has several limitations. First, the datasets used in this study are only in English. An analysis of the viability of the method on multi-lingual data needs to be investigated to ensure broad applicability. Second, our experiments were limited to three datasets for stance detection. The work can be expanded to other datasets for stance detection as well as other tasks in argument mining. Third, we only evaluated five large language models. The investigation could be expanded to other LLMs that are available in multiple sizes as we note that they suffer from a significant bias towards the support label, which requires further testing. Fourth, due to computational constraints, our experiments were conducted as a single run using Greedy Decoding. While this ensures deterministic outputs, future work should explore multi-run evaluations across varying generation temperatures and example selection methods to assess model stability.

Acknowledgments

Oana Cocarascu was partially funded by EPSRC (grant number EP/X04162X/1).

References

Yamen Ajjour and Henning Wachsmuth. 2025. Exploring LLM priming strategies for few-shot stance classification. In *Proceedings of the 12th Workshop on Argument Mining*. Association for Computational Linguistics.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Seth Aycock and Rachel Bawden. 2024. Topic-guided example selection for domain adaptation in LLM-based machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017a. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. 2017b. Improving claim stance classification with lexical knowledge expansion and context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark. Association for Computational Linguistics.

Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2025. Can large language models understand argument schemes? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13666–13681, Vienna, Austria. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. *Preprint*, arXiv:1906.03538.

Oana Cocarascu, Elena Cabrio, Serena Villata, and Francesca Toni. 2020. A dataset independent set of baselines for relation prediction in argument mining. *Preprint*, arXiv:2003.04970.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2022. Don't drop the topic - the role of the prompt in argument identification in student writing. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 124–133, Seattle, Washington. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy. Association for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2024. Can large language models perform relation-based argument mining? *Preprint*, arXiv:2402.11243.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can large language models perform relation-based argument mining? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534, Abu Dhabi, UAE. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qinyu Han, Zhihao Yang, Hongfei Lin, and Tian Qin. 2024. Let topic flow: A unified topic-guided segment-wise dialogue summarization framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2021–2032.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Patrick Jansson and Shuhua Liu. 2017. Distributed representation, LDA topic modelling and deep learning for emerging named entity recognition from social media. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 154–159, Copenhagen, Denmark. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Omid Kashefi, Sophia Chan, and Swapna Somasundaran. 2023. Argument detection in student essays under resource constraints. In *Proceedings of the 10th Workshop on Argument Mining*, pages 64–75, Singapore. Association for Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia Est—On the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10:1–10:25.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, and 30 others. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

OpenAI. Introducing chatgpt.

Yasser Otiefy and Alaa Alhamzeh. 2024. Exploring large language models in financial argument relation identification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 119–129, Torino, Italia. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2018. Is something better than nothing? automatically predicting stance-based arguments using deep learning and small labelled dataset. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 28–34, New Orleans, Louisiana. Association for Computational Linguistics.

Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021a. No NLP task should be an island: Multi-disciplinarity for diversity in news recommender systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 45–55, Online. Association for Computational Linguistics.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021b. Is stance detection topic-independent and cross-topic generalizable? - a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Ramon Ruiz-Dolz and Javier Iranzo-Sánchez. 2024. Vivesdebate-speech: A corpus of spoken argumentation to leverage audio features for argument mining. *Preprint*, arXiv:2302.12584.

Azam Seilsepour, Reza Ravanmehr, and Ramin Nassiri. 2023. Topic sentiment analysis based on deep neural network using document embedding technique. *J. Supercomput.*, 79(17):19809–19847.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Qingfa Xiao, Shuangyin Li, and Lei Chen. 2023. Topic-dpr: Topic-based prompts for dense passage retrieval. *Preprint*, arXiv:2310.06626.

Jiaqing Yuan, Ruijie Xi, and Munindar P. Singh. 2024. A benchmark for cross-domain argumentative stance classification on social media. *Preprint*, arXiv:2410.08900.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Topic-Guided Prompts

In this section, we provide examples of the exact prompts generated by our Topic-Guided method. As described in Section 3, each prompt begins with in-context examples specifically retrieved based on the topic of the target instance. These demonstration examples consist of two supporting and two opposing arguments. Following the examples, the model is provided with task instructions and the target pair of arguments to classify. Table 5 illustrates a prompt structure for the IBM dataset, Table 6 shows a prompt for the Perspectrum dataset, and Table 7 demonstrates a prompt for the UKP dataset.

Prompt for the topic "the monarchy"
Arg1: This house would abolish the monarchy Arg2: a morally-based, balanced monarchy is stressed as the ideal form of government Relation: attack Arg1: This house would abolish the monarchy Arg2: The system of monarchy since antiquity has contrasted with forms of democracy Relation: support Arg1: This house would abolish the monarchy Arg2: The hereditary nature of the monarchy is said to conflict with egalitarianism and dislike of inherited privilege Relation: support Arg1: This house would abolish the monarchy Arg2: The monarch serves as a ceremonial figurehead symbol of national unity and state continuity Relation: attack In this task you will be given two arguments and your goal is to classify the relation between them as either 'support' or 'attack' based on the definitions below. Support: It is an argument that is in favour of the parent argument. Attack: It is an argument that contradicts or opposes the parent argument. Complete this new example with the relation between the two arguments, support or attack. Answer with ONLY one word.
New pair of arguments
Arg1: This house would abolish the monarchy Arg2: The principal advantage of hereditary monarchy is the immediate continuity of leadership Relation:

Table 5: Example of a prompt used when classifying the stance of arguments from the IBM dataset, for the topic "the monarchy".

Prompt for the topic "environment"

Arg1: Build hydroelectric dams

Arg2: Hydroelectric dams cause serious problems in our communities .

Relation: attack

Arg1: Any vaccines should be required for children

Arg2: Vaccines are economically efficient, versus the cost of disease.

Relation: support

Arg1: Co2 Does Not Cause Global Warming

Arg2: There is no evidence that Global warming is accurate as it is based on computer modelling.

Relation: support

Arg1: Co2 Does Not Cause Global Warming

Arg2: The climate is changing and people are dying.

Relation: attack

In this task you will be given two arguments and your goal is to classify the relation between them as either 'support' or 'attack' based on the definitions below.

Support: It is an argument that is in favour of the parent argument.

Attack: It is an argument that contradicts or opposes the parent argument.

Complete this new example with the relation between the two arguments, support or attack. Answer with ONLY one word.

New pair of arguments

Arg1: Antarctica Should Be Opened Up For Resource Exploitation

Arg2: Antartic resource exploitation is unevitable: Let's do it orderly!

Relation:

Table 6: Example of a prompt used when classifying the stance of arguments from the Perspectrum dataset, for the topic "environment".

Prompt for the topic "minimum wage"

Arg1: minimum wage is good.

Arg2: not true : the typical minimum wage worker is not a high school student earning weekend pocket money .

Relation: attack

Arg1: minimum wage is good.

Arg2: milton friedman called them a form of discrimination against low-skilled workers .

Relation: support

Arg1: minimum wage is good.

Arg2: it would be a move in the direction of the wages of decent living - a performance , one might say , of decency itself .

Relation: support

Arg1: minimum wage is good.

Arg2: and those employers , in turn , would be unable to hire as many people an undesirable result when unemployment continues to hover at about 8 percent .

Relation: attack

In this task you will be given two arguments and your goal is to classify the relation between them as either 'support' or 'attack' based on the definitions below.

Support: It is an argument that is in favour of the parent argument.

Attack: It is an argument that contradicts or opposes the parent argument.

Complete this new example with the relation between the two arguments, support or attack. Answer with ONLY one word.

New pair of arguments

Arg1: minimum wage is good.

Arg2: raising the minimum wage does more harm than good .

Relation:

Table 7: Example of a prompt used when classifying the stance of arguments from the UKP dataset, for the topic "minimum wage".

AMResources: Cataloging Argument Mining Datasets

Dexter Williams[1], Shiwei Liu[4], Manfred Stede[2], Henning Wachsmuth[3], Jodi Schneider[1]

[1]University of Wisconsin-Madison, [2]Potsdam University, [3]Leibniz University Hannover, [4]University of Illinois Urbana-Champaign

Correspondence: djwilliams22@wisc.edu and jschneider@pobox.com/jodi@wisc.edu

Abstract

Annotated datasets are essential for developing and evaluating argument mining systems, yet information about argument mining datasets remains scattered across papers, repositories, and task-specific surveys. To address this, we introduce *AMResources* (`http://purl.archive.org/amresources`), an online catalog that organizes argument mining datasets by task, and captures relationships among datasets, releases, and papers. We draw particular attention to relationships such as re-annotation and dataset extension. To curate dataset information into a consistent and provenance-aware structure, *AMResources* links datasets to canonical papers. For each dataset release, *AMResources* records standardized metadata such as language, genre, unit type and unit count, annotator characteristics, agreement reporting, and accessibility. We argue that such structured dataset documentation remains critical in the era of large language models, where annotated datasets increasingly serve as high-quality evaluation benchmarks and where tracing dataset provenance and annotation layers is necessary for systematic comparisons across tasks.

1 Introduction

Datasets are an essential resource for argument mining tasks, supporting both the training of computational models and the evaluation of automated methods. However, a number of factors (theoretical and practical) affect the cataloging of argument mining datasets. Argument mining datasets may be annotated for a number of different tasks and subsequently annotated for a different purpose (Stab and Gurevych, 2017a; Goffredo et al., 2023). Annotated datasets differ in the quality of annotators used, and may or may not report agreement. Datasets vary in their availability, size, language, and format. A dataset may also stand in an important relationship to other datasets, through the inheritance of previous layers of annotation or subsets of units.

Despite these challenges, argument mining datasets have been cataloged and discussed by the community. Habernal and Gurevych (2017) provide a table of 23 previous works on annotating argumentation, documenting the properties of created datasets. A chapter of *Argumentation Mining* (Stede and Schneider, 2018) examines several annotation schemes and catalogs 10 relevant datasets. Janier and Saint-Dizier (2019) includes a chapter on argument annotation, covering annotation schemes, tools, and 8 relevant datasets. Lawrence and Reed (2019) discusses a number of datasets, focusing on the challenges of producing and annotating datasets for argument mining.

A number of online sites have been created to make argument mining-related resources more findable, although their scope and maintenance varies. Persiani et al. (2024) provides a website[1] containing an ongoing survey of tools useful for computational argumentation; the site is frequently updated and accepts content requests from the community through a GitHub repository.[2] Guerraoui et al. (2023) surveys NLP feedback systems in argumentation; a supporting website[3] categorizes works on datasets, tooling, and computational models into four sections (richness, visualization, interactivity, and personalization).

A few sites catalog argument mining datasets. The Ubiquitous Knowledge Processing (UKP) Lab at TU Darmstadt[4] and the Webis Group[5] publish datasets used in their argument mining research.

[1]`https://people.cs.umu.se/~tkampik/argtools/page/index.html`

[2]`https://github.com/TimKam/fantastic-arg-tools`

[3]`https://kmilia.github.io/teach_me_how_to_argue/`

[4]`https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/1359`

[5]`https://webis.de/data.html`

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 52–58
July 3, 2026

AIFdb (Lawrence and Reed, 2014; Lawrence et al., 2016) is a large community-sourced database of argument mining datasets.[6] Datasets in *AIFdb* are stored in the Argument Interchange Format (Chesñevar et al., 2006), which enables interoperability with argument mapping tools such as *OVA+* (Janier et al., 2014). However, *AIFdb* does not enforce documentation of dataset properties. This makes it challenging to use *AIFdb* to search for datasets for a specific argument mining task. In contrast, the documentation of dataset properties is more evident in the website introduced by Romberg et al. (2025) to accompany their survey of datasets for automatic argument quality assessment.[7] Their website categorizes argument quality datasets according to a taxonomy, and lists a number of properties of each dataset such as annotator agreement, annotator type, source, and whether the dataset is an extension of another dataset.

In this paper we introduce an online catalog of argument mining datasets: *AMResources*.[8] In contrast to other sites, *AMResources* emphasizes the suitability of datasets for specific argument mining tasks, and makes dataset links findable. *AMResources* catalogs argument mining datasets across a plethora of argument mining tasks through a linked structure of *datasets*, *releases*, *papers*, and *annotations*. Our work is inspired by the catalog of Romberg et al. (2025). During the creation of Stede et al. (2026), the second edition of (Stede and Schneider, 2018), it became apparent that the rapid growth of argument mining datasets could not be covered in a single chapter; this led to the development of our online catalog as a maintainable resource for discovering argument mining datasets. The initial release of *AMResources* contains the 39 datasets described in the 'Selected Corpora' appendix of Stede et al. (2026).

2 Connecting Datasets, Papers, and Annotations

We consider a set of units (documents, tweets, etc.) to be a *dataset*. A dataset has a number of properties such as a *domain*, *unit count*, and *unit type*. The first publication of a *dataset* is the *source dataset*. Any newer dataset which was published after the source dataset, with units that are a subset of the units of the source dataset, is a *release* of the source dataset.

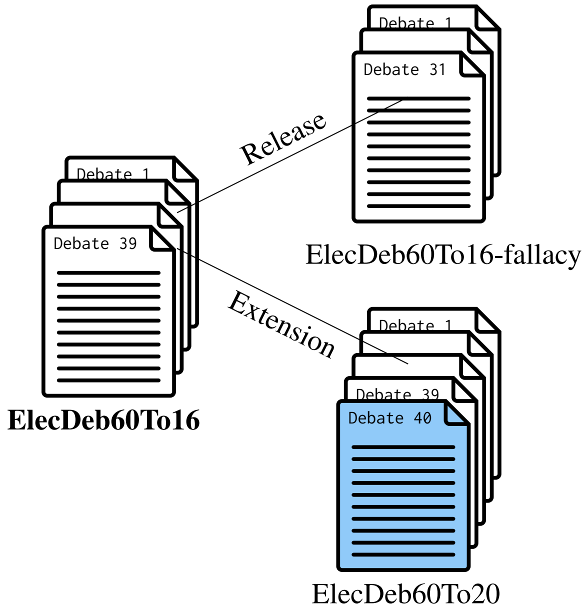


Figure 1: Release and extension relationships between the ElecDeb series of datasets. ElecDeb60To16 is a source dataset. ElecDeb60To16-fallacy is a release of ElecDeb60To16 as it consists of a subset of debates from ElecDeb60To16. ElecDeb60To20 extends ElecDeb60To16 with a new debate transcript.

Releases serve an important function in capturing inheritance relationships between datasets. The units within one dataset might just be a subset of the units from another dataset. Frequently, datasets are annotated multiple times, with each annotated dataset being given a new name. For example the ElecDeb60To16 dataset (Haddadan et al., 2019), a dataset of political transcripts from 1960 to 2016, was subsequently annotated and released as the ElecDeb60To16-fallacy dataset (Goffredo et al., 2022). Thus, ElecDeb60To16 is the source dataset, with a later release in ElecDeb60To16-fallacy.

It is also possible for a newer dataset to include a subset of the units of a source dataset while also adding new units. In this case we do not call the newer dataset a release of the source dataset, but rather an *extension* of the source dataset. For example the ElecDeb60To20 dataset (Goffredo et al., 2023) contains an additional four years of political transcripts from 2016 to 2020 as well as the original political transcripts in ElecDeb60To16. Therefore we call ElecDeb60To20 an *extension* of ElecDeb60To16. Relationships between the three ElecDeb datasets can be seen in Figure 1.

Our approach allows us to capture the evolution

[6] https://www.aifdb.org/search

[7] https://goofy-grouse-1da.notion.site/Database-e3e5886191ef472aaffb47fec0daea92

[8] http://purl.archive.org/amresources

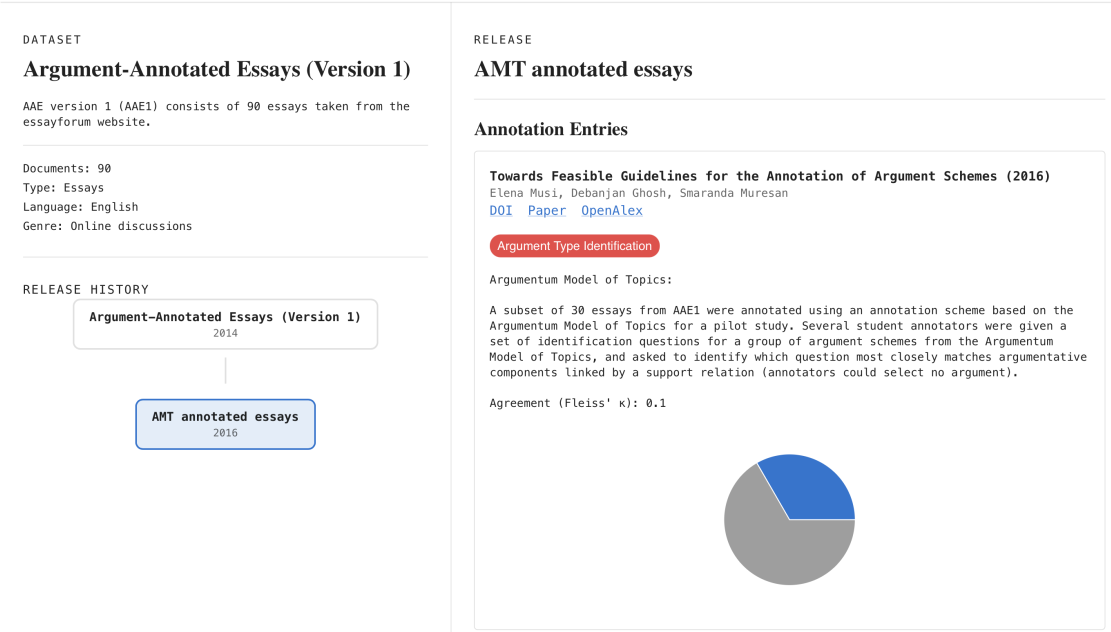


Figure 2: Example view from *AMResources*. The user has selected details on AMT annotated essays (Musi et al., 2016) which is a release of the Argument Annotated Essays Corpus (Stab and Gurevych, 2014). The left pane displays other releases for the dataset. The right pane shows details on the release, including a pie chart showing how much of the original dataset was annotated in the release.

of properties of argument mining datasets, instead of presenting a flat table of datasets. Datasets that have been annotated multiple times in different releases are particularly valuable to identify. Multi-layered argument annotations on a dataset can be used to identify dependencies between different argumentative relations (Chistova, 2023). Some argument mining datasets have been annotated repeatedly, for instance the Argument-Annotated Essays Corpus has been annotated in at least eight different papers (Carlile et al., 2018; Stab and Gurevych, 2014, 2016, 2017a,b; Musi et al., 2016; Schaefer et al., 2023; Marro et al., 2022).

Finally, a *paper* is an academic work that either introduces a source dataset or annotates an existing dataset to produce a new release of that dataset. An *annotation* is carried out by a paper on a dataset, to support a particular argument mining task (see Section 3). *Datasets*, *releases*, *papers*, and *annotations* form the four linked objects of *AMResources*.

3 Argument Mining Tasks

Beyond the surface-level properties of datasets, we aim to categorize argument mining datasets by their potential use. In particular, we map each dataset to the set of argument mining tasks that it enables studying. We differentiate between two broad categories of argument mining tasks based on their use of argumentative discourse units (ADUs): *ADU-concerned* and *ADU-agnostic*.

ADU-Concerned Tasks This category covers tasks that, in some way, deal with the collection, classification, and linking of ADUs. Our selection of ADU-concerned tasks is taken from Stede et al. (2026, Ch. 4) and forms a classic argument mining pipeline. The description of ADU-concerned tasks is shown in Table 1.

ADU-Agnostic Tasks These tasks do not necessarily involve ADUs but undebatably fall under the notion of argument mining in the broader sense (see Table 2). For instance, argument quality assessment might not involve annotating argument structure, and instead assigning binary scores of quality to argumentative text (Toledo et al., 2019). We distinguish between minimal and maximal argument quality assessment, a distinction taken from Wachsmuth et al. (2024). Another ADU-agnostic task is to mine the full type/form/scheme of an argument from text (e.g., Schneider et al., 2013); we call this argument type identification.

4 AMResources: A Website for Argument Mining Datasets

We host our collection of datasets on a website called *AMResources*, deployed as a *GitHub Pages*

Task	Description
Component Segmentation	Distinguishes argumentative from non-argumentative text portions and demarcates individual argument components
Component Type Classification	Determines the types of given argument components, using a coarse-grained (e.g., claim, premise) or a more fine-grained (e.g., policy claim, fact claim, value claim) type inventory
Relation Identification	Detects the presence of some (indeterminate) argument relation between two argument components
Relation Type Classification	Determines the types of given relations (e.g., support, attack)

Table 1: ADU-concerned tasks: The four tasks that deal with the collection, classification, and linking of ADUs, from Stede et al. (2026, Ch. 4).

Task	Description
Maximal Quality Assessment	Assesses the quality of an argument based on a theory that describes what an argument should *ideally* be
Minimal Quality Assessment	Assesses the quality of an argument based on a theory that describes what an argument should *avoid* being
Type Identification	Identifying the type/form of the whole argument

Table 2: Three examples of ADU-agnostic tasks, from Wachsmuth et al. (2024) and (Schneider et al., 2013).

site.[9] The repository for the site is public, and we encourage contributions of new datasets, and include a form for requests for new datasets to be added. We store all covered dataset metadata in two JSON files (one for dataset metadata and one for paper metadata) which can be downloaded through the GitHub repository.

AMResources presents releases in a searchable table view (release, dataset, language, genre, unit type, unit count, and task tags). The catalog can be exported as a CSV file. Selecting a row opens a detail view (see Figure 2) with dataset metadata, a release history visualization, and structured annotation summaries linked to papers. Where possible we provide a DOI and an OpenAlex link for papers. Each annotation entry contains a detailed summary, and records task labels, annotator type, agreement reporting, accessibility, and a release link.

5 The Future Role of Argument Mining Datasets

Argument mining datasets have been, and continue to be, an important artifact of argument mining research. However, it is prudent to reflect on the role that argument mining datasets now serve in light of advancements in large language models (LLMs). The rise in the use of LLMs reflects a shift away from classical supervised learning methods which relied on annotated datasets, towards methods which involve little to no need for manually annotated data for training.

Despite a decreasing relevance to model training, argument mining datasets continue to be important for evaluation. Several argument mining tasks (such as ADU segmentation) are not straightforward to model with LLMs, leading to a demand for LLM-oriented argument mining benchmarks (Dhole et al., 2025; Gemechu et al., 2024; Gurjar et al., 2025; Ajjour et al., 2026).

Additionally, annotated datasets in argument mining also have a unique role in comparison to other fields. Argument mining has an unusually strong theoretical foundation compared to many other NLP tasks. This heavy emphasis on theory is manifested in the annotations, and it has the potential to inform NLP methods that go beyond pure statistics (Lauscher et al., 2022; Wachsmuth et al., 2024). In particular, competing theoretical models of argument underlie argument mining. These competing theoretical models make their own assumptions and simplifications (Cardoso et al., 2023). Datasets with multiple layers of annotation provide an opportunity to study dependencies between argument mining tasks and argument models. In this respect, annotated argument mining datasets can help us to inform our theoretical models of arguments, and how these models interface with text.

[9] http://purl.archive.org/amresources

6 Conclusion

We presented AMResources, a web catalog for argument mining datasets that represents provenance through explicit links among datasets, releases, and papers. We argue that such structured dataset documentation remains critical in the era of large language models, where annotated datasets increasingly serve as high-quality evaluation benchmarks and where tracing dataset provenance and annotation layers is necessary for systematic comparisons across tasks. Going forward, we will expand coverage through community contributions in order to keep *AMResources* up-to-date over time.

Acknowledgments

Dexter Williams was supported by University of Wisconsin-Madison startup funds to Jodi Schneider as well as her University of Illinois Campus Research Board RB25118: Support for a Second Edition of the Argumentation Mining monograph, which also supported work from Shiwei Liu. Some support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation. Jodi Schneider was partially funded as the 2024-2025 Perrin Moorhead Grayson and Bruns Grayson Fellow at Harvard University's Radcliffe Institute for Advanced Study. Henning Wachsmuth is partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/3 2026 – 438445824.

CRediT statement

Dexter Williams: Writing - Original Draft, Data Curation, Software, Project Administration
Shiwei Liu: Data Curation, Writing - Review & Editing
Manfred Stede: Conceptualization, Writing - Review & Editing
Henning Wachsmuth: Conceptualization, Writing - Review & Editing
Jodi Schneider: Conceptualization, Supervision, Funding Acquisition, Writing - Review & Editing
Markus Rutledge: Data Curation

References

Yamen Ajjour, Carlotta Quensel, Nedim Lipka, and Henning Wachsmuth. 2026. ArgBench: Benchmarking LLMs on computational argumentation tasks. *Preprint*, arXiv:2604.17366.

Henrique Lopes Cardoso, Rui Sousa-Silva, Paula Carvalho, and Bruno Martins. 2023. Argumentation models and their use in corpus annotation: Practice, prospects, and challenges. *Natural Language Engineering*, 29(4):1150–1187.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Carlos Chesñevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerard Vreeswijk, and Steven Willmott. 2006. Towards an Argument Interchange Format. *The Knowledge Engineering Review*, 21(4):293–316.

Elena Chistova. 2023. End-to-end argument mining over varying rhetorical structures. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3376–3391, Toronto, Canada. Association for Computational Linguistics.

Kaustubh Dhole, Kai Shu, and Eugene Agichtein. 2025. ConQRet: A new benchmark for fine-grained automatic evaluation of retrieval augmented computational argumentation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5687–5713, Albuquerque, New Mexico. Association for Computational Linguistics.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. ARIES: A general benchmark for argument relation identification. In *Proceedings of the 11th Workshop on Argument Mining*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence Main Track*, volume 5, pages 4143–4149, Vienna, Austria. International Joint Conferences on Artificial Intelligence.

Camelia Guerraoui, Paul Reisert, Naoya Inoue, Farjana Sultana Mim, Keshav Singh, Jungmin Choi, Irfan Robbani, Shoichi Naito, Wenzhi Wang, and Kentaro Inui. 2023. Teach me how to argue: A survey on NLP feedback systems in argumentation. In *Proceedings of the 10th Workshop on Argument Mining*, pages 19–34, Singapore. Association for Computational Linguistics.

Omkar Gurjar, Agam Goyal, and Eshwar Chandrasekharan. 2025. ArgCMV: An argument summarization benchmark for the LLM-era. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21870–21883, Suzhou, China. Association for Computational Linguistics.

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

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! Mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Mathilde Janier, John Lawrence, and Chris Reed. 2014. OVA+: An argument analysis interface. In *Computational Models of Argument*, pages 463–464. IOS Press.

Mathilde Janier and Patrick Saint-Dizier. 2019. *Argument mining: Linguistic foundations*. Wiley-ISTE, Hoboken.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—On the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

John Lawrence, Mathilde Janier, and Chris Reed. 2016. Working with open argument corpora. In *Argumentation and Reasoned Action: Proceedings of the First European Conference on Argumentation*, pages 367–380, Lisbon, Portugal. College Publications.

John Lawrence and Chris Reed. 2014. AIFdb corpora. In *Computational Models of Argument*, pages 465–466. IOS Press.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Santiago Marro, Elena Cabrio, and Serena Villata. 2022. Graph embeddings for argumentation quality assessment. In *Findings of the Association for Computational Linguistics*, pages 4154–4164, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2016. Towards feasible guidelines for the annotation of argument schemes. In *Proceedings of the Third Workshop on Argument Mining*, pages 82–93, Berlin, Germany. Association for Computational Linguistics.

Michele Persiani, Esteban Guerrero, Andreas Brännström, Kaan Kilic, and Timotheus Kampik. 2024. Fantastic argumentation tools and where to find them. In *Proceedings of the Fifth International Workshop on Systems and Algorithms for Formal Argumentation co-located with 10th International Conference on Computational Models of Argument, Hagen, Germany, September 17th, 2024*, CEUR Workshop Proceedings, pages 56–68. CEUR-WS.org.

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. 2025. Towards a perspectivist turn in argument quality assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7458–7485, Albuquerque, New Mexico. Association for Computational Linguistics.

Robin Schaefer, René Knaebel, and Manfred Stede. 2023. Towards fine-grained argumentation strategy analysis in persuasive essays. In *Proceedings of the 10th Workshop on Argument Mining*, pages 76–88, Singapore. Association for Computational Linguistics.

Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 conference on Computer Supported Cooperative Work*, CSCW '13, pages 1069–1080, San Antonio, Texas, USA. Association for Computing Machinery.

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, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining*, pages 113–118, Berlin, Germany. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab and Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990, Valencia, Spain. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation mining*. Number 40 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Manfred Stede, Jodi Schneider, and Henning Wachsmuth. 2026. *Argument mining*, second edition. Springer Nature. Forthcoming.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - New datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, and Timon Ziegenbein. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 1519–1538, Torino, Italia. ELRA and ICCL.

Argument-Based Comparative Question Answering Evaluation Benchmark

Irina Nikishina[1] **Saba Anwar**[1] **Nikolay Dolgov**[2]
Maria Manina[2] **Daria Ignatenko**[2] **Viktor Moskvoretskii**[2,4]
Artem Shelmanov[3] **Tim Baldwin**[3] **Chris Biemann**[1]

[1]University of Hamburg [2]HSE University [3]MBZUAI [4]Skoltech
Correspondence: irina.nikishina@uni-hamburg.de

Abstract

Despite the ability of large language models (LLMs) to generate coherent comparative answers, automatic comparative question answering (CQA) remains challenging due to the absence of standardized evaluation criteria and the high resource demands of manual assessment. To address these problems, this paper proposes a comprehensive evaluation framework designed to assess the quality of CQA summaries using LLMs-as-a-Judge. We formulate 15 evaluation criteria for assessing comparative answers generated by various sources, including LLMs, human experts, and prior work. To capture a diverse range of comparative answers, LLM summaries were generated under various prompting scenarios. We evaluate the effectiveness of our framework using both human assessment and LLMs, demonstrating the consistency between automated and manual evaluations. Finally, we fine-tune Llama-3-8B-Instruct on a dataset generated from the best-performing CQA models in our evaluation benchmark to assess whether open-source models can replace closed-source LLMs. All used data, code, and evaluation results are publicly available[1].

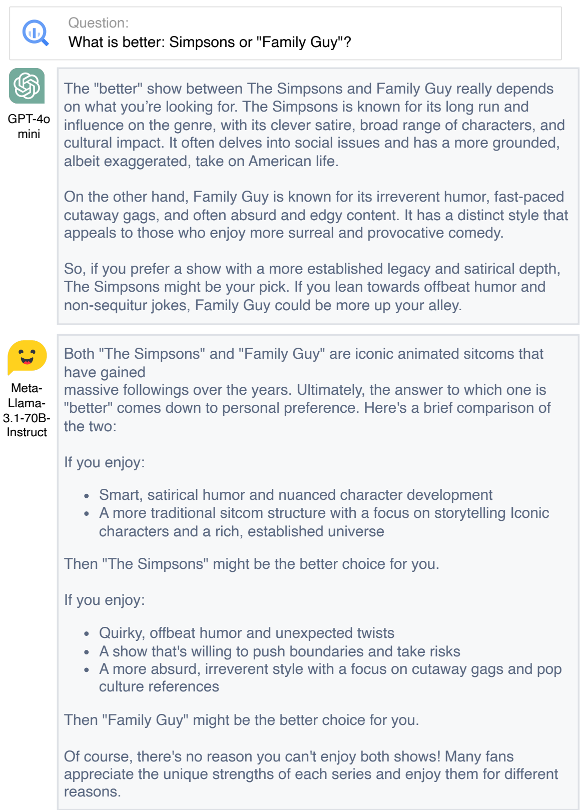


Figure 1: Examples of LLM answers to comparative questions (shortened).

1 Introduction

Are *"Nike"* sneakers better than *"Adidas"*? Is an *"iPhone"* easier to use than an *"Android"* phone? What is faster, *"Windows"* or *"Linux"*? Such questions are common user requests on the web. Making an informed choice requires evaluating and comparing various characteristics of the options under consideration. For instance, when buying a car, it is important to consider the brand reputation, the fuel efficiency, comfort, and general performance. Conducting thorough comparisons manually by scanning different websites requires time and effort, especially given the vast amount of data available online (Iso et al., 2022). In such cases, a comparative answer or summary can serve as an effective and time-saving solution.

Comparative Question Answering (CQA) is a Natural Language Processing task that aims to answer requests containing two or more objects by choosing one of them and supporting the answer with arguments. The final goal is not necessarily to come to a distinct final choice, but to describe each object's advantages and disadvantages, taking into account the specified aspect(s). This task can also be viewed as an abstractive comparative summarization. In this paper, we focus on comparing only two objects, leaving the cases with three or more

[1]https://github.com/lilaspourpre/cqa-evaluation-framework/README.md

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 59–71
July 3, 2026 ©2026 Association for Computational Linguistics

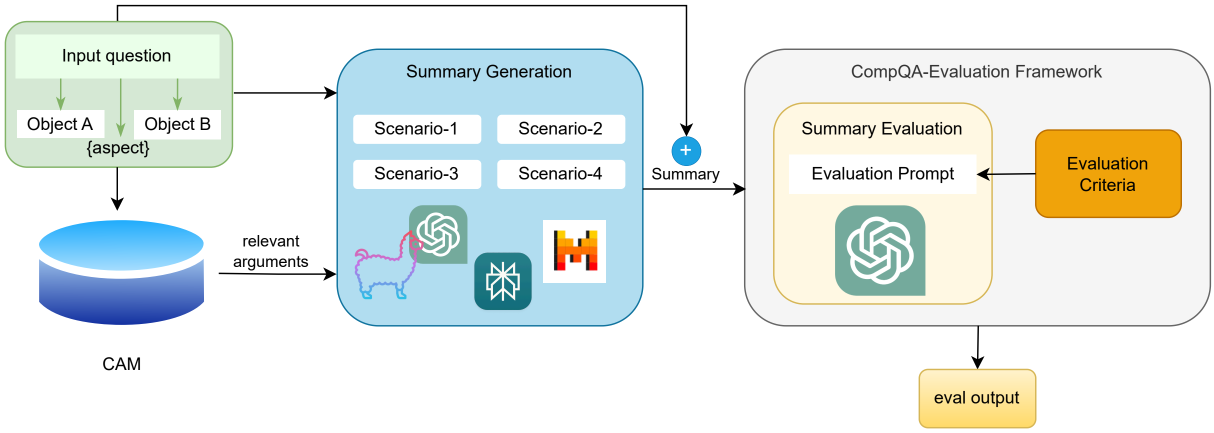


Figure 2: Overall pipeline of the CompQA evaluation dataset creation and evaluation framework.

for future research. Unlike traditional opinion summarization, which focuses on aggregating views about a single object, the task of generating comparative summaries is more complex, as it involves evaluating each object in the context of the other. Traditional approaches to this task focus only on the initial steps of CQA that precede comparative answer generation and only partially address the Argument Summarization task (Chekalina et al., 2021; Shallouf et al., 2024) for answer generation.

Large Language Models (LLMs) have emerged as powerful tools for a wide range of tasks, including search and recommendations. They significantly simplify the process of generating comparative answers, often providing coherent and well-structured responses (see example in Figure 1). However, the quality of these responses varies greatly, posing a persistent challenge in the automated evaluation of such summaries. Prior work suggests that a well-constructed abstractive summary should be *coherent*, *concise*, *factually consistent*, *relevant*, *non-redundant*, *grammatically accurate*, and exhibit *high readability* (Gupta and Gupta, 2019; Kryscinski et al., 2019; Shakil et al., 2024). Achieving this type of qualitative assessment often requires human evaluation, guided by a structured framework that decomposes the evaluation into specific criteria. Importantly, this framework may vary depending on the nature of the summary: for instance, summarizing a Wikipedia article requires different criteria (e.g. factual sequences) than summarizing a comparison between two products. Consequently, there is no consensus regarding the actual criteria a summary must follow. Moreover, manual evaluation is a tedious task; therefore, its automation is highly beneficial for the research and development of CQA systems.

In this paper, we aim to fill the mentioned gaps in benchmarking the CQA systems and answer the following research questions: *(RQ1) Which criteria should be considered for assessing a comparative summary?*, *(RQ2) Can LLMs reliably evaluate comparative summaries with human expert-level quality?*, *(RQ3) How do different LLMs fare against each other in generating high-quality summaries?*

The contributions of this work are as follows:

1. Systemizing previous efforts in CQA, we develop 15 criteria for scoring comparative answers and implement automatic CompQA evaluation pipeline based on a LLM-as-a-Judge approach (Figure 2) that is able to assess CQA summaries on the basis of these criteria. We verify the usefulness of the proposed benchmark by comparing automatic assessments with human judgements. As a part of benchmark, we create two datasets of 50 (main) and 432 (extended) comparative questions with object pairs, aspects, and arguments for model evaluation.

2. Using the developed benchmark, we conduct the first automatic and manual evaluation of several LLMs on the CQA task. We utilize GPT-3.5, GPT-4, Llama-3-8B-Instruct, Llama-3-70B-Instruct, Perplexity, and Mixtral for the main experiments, while Gemma-3-27B, Qwen3-8B, DeepSeek-R1, and Gemini-2.5-Flash are used for additional experiments. To ensure a comprehensive performance evaluation, we use CQA datasets sourced from (Nikishina et al., 2025), CAM 2.0 (Shallouf et al., 2024), and Yahoo! Answers (Chekalina et al., 2021), along with comparative answers gener-

ated by a subset of these LLMs under various prompting scenarios

3. We demonstrate that the performance gap between open- and closed-source models can be narrowed by fine-tuning Llama-3-8B-Instruct on a high-quality dataset distilled from our evaluation benchmark.

All code, datasets, and evaluation results are available online[1].

Characterstics	
Total Pairs	50
Pairs with defined aspect	20
Average # of arguments	10
Min # of arguments	3
Max # of arguments	20
Pairs with more than average # of arguments	23

Table 1: Statistics of data retrieved from (Chekalina et al., 2021) and CAM 2.0 (Shallouf et al., 2024).

2 CompQA Evaluation Framework

Defining criteria for a high-quality answer is crucial, as checking only for coherence and logic—or using metrics like word and vector similarity—is not enough. This is especially relevant to our usecase: when generating answers, we use a list of arguments, and the output is directly tied to the complexity and level of detail of the input prompt (Loya et al., 2023). Thus, the generated answer can vary from an unstructured paragraph to a well-organized summary with clearly defined components, depending entirely on the input. This also means that two summaries comparing the same pair of objects can have a low token-match metric if different arguments were used, despite the fact that both may still be high-quality comparative texts.

Assuming two objects are to be compared, we define a good-quality comparative answer as one that helps the user *decide* between the objects by comparing several *relevant aspects*. These aspects can be general or specified by the user. In establishing our criteria for such an answer, we adopt metrics that align with human perception of quality, requiring that a summary should be *well-structured*, *concise*, *factual*, *relevant*, *coherent*, *informative*, and *non-redundant*. Most of these criteria are primarily adapted from the work of Peyrard (2019) and Kryscinski et al. (2019), as well as recent findings on LLM-based human-like evaluation (Gao et al., 2023; Liu et al., 2023). Thus, to answer the **RQ-1** on developing criteria for the comparative question-answer summary, we define and categorize the following dimensions to assess generated answers:

1. *structure*: a well-defined structure that has: a) a short introductory summary, b) a list of *named aspects* with *short descriptions*, and c) a distinct *final choice*.

2. *relevancy*: arguments are relevant to the aspect of comparison (if given), compare both objects, and are ordered from more generic to specific ones or vice versa.

3. *quality*: should further fulfill the following aspects: *concise* (optimal length focusing on key information, should be between 12 to 20 sentences, aligning with the dataset's argument distribution to ensure comprehensive yet brief coverage), *factual* (no hallucination or contradictory arguments), *informative* (if the summary is required for a specific aspect, then the introduction, arguments, and conclusion should incorporate that, otherwise they should be generic), *coherence* (comparison logic is easy to follow, statements are not self-contradictory or repetitive, the conclusion does not contradict the arguments).

2.1 Automatic Evaluation using LLM-as-a-Judge

In this section, we present an evaluation framework called CompQA that implements these quality checks with 15 questions and assigns a score between 0 and 19 points. Each question assigns a maximum score of 1 or 2, depending on various scenarios. The criteria are developed to simplify the annotation and increase inter-annotation agreement of human annotators. Figure 3 presents the CompQA framework and explains each scored point for each criterion in detail.

To automatically score the CQA answers, we prompt an LLM. The template is presented in Figure 5. Because of appendix size limitation, the complete prompt scenario can be found in the Github repository[1]. The output of the model is expected to be a JSON dictionary where the keys are criteria numbers and values are the scores assigned by LLM. After preliminary experiments, we have decided not to require models to provide the total score, as the total value is usually not the same as the actual sum. The model provides the scores according to each criterion.

structure / 7	a short introduction is present	1	- the introduction is missing or is too long - 0 points - the introduction is short and concise - 1 point
	defined aspects used for comparison in the whole comparison	1	- the comparison is arbitrary with no specific aspects - 0 points - the comparison uses specific aspects to compare objects - 1 point
	the introduction mentions the most important comparison aspects	1	- no aspects are mentioned or no introduction - 0 points - several most important aspects are mentioned in the introduction - 1 point
	the main body of comparison has good structure	1	- some aspects mix with others, the structure is harder to follow - 0 point - the aspects are logically divided into separate aspects - 1 point
	the main body of the comparison has defined aspect names	1	- no aspect names are given, comparison is inconcrete - 0 points - main body has distinct aspect names - 1 point
	the main body of the comparison has defined aspect descriptions	1	- no aspect descriptions are given, comparison is inconcrete - 0 points - main body has distinct aspect descriptions - 1 point
	the final choice is given explicitly	1	- no explicit choice made or lengthy justification present - 0 points - short and explicit choice made - 1 point
relevance / 5	the comparison aspects in the main body of the comparison are sorted by general applicability	1	- statements are not sorted at all - 0 points - statements are sorted by general/important statements first, specific statements closer to the end - 1 point
	each argument is relevant to the aspect of comparison (if any, otherwise is general and is not biased towards any aspect)	2	- most arguments are irrelevant - 0 points - most arguments are relevant - 1 point - all arguments are relevant - 2 points
	each argument compares both objects evaluating each argument separately & scale up to 2	2	- some arguments do not compare the objects - 0 points - some arguments give information only about one object - 1 point - all arguments compare both objects - 2 points
quality / 7	there are no hallucinations or statements contradicting common knowledge	2	- many hallucinations, serious factual inaccuracy - 0 points - some hallucinations, but mostly correct - 1 point - no hallucinations, factually correct - 2 points
	the comparison has proper language and is easy to follow	2	- hard to read, profanity present or illogical - 0 points - some grammar issues, broken logic - 1 point - no grammar issues, good structure and logic - 2 points
	there are no repetitive statements or statements too similar to each other	1	- some statements repeat others' meaning very closely - 0 points - all statements are unique and do not repeat - 1 points
	the final answer is concluded from the statements in the main body and takes the main aspect (if there is one) into consideration	1	(if all statements favor object 1, then the answer is object 1 if both objects are equally good or equally bad, then none of the objects is preferred and the answer is inconclusive) - the final answer is not concluded from the arguments or main aspect (if there is one) or no answer is given - 0 points - the final answer is concluded from the majority of arguments and main aspect (if there is one) - 1 point
	the summary itself is not too short and not too long	1	- the summary is too short (less than 12 sentences) or too long (more than 20 sentences) - 0 points - the summary is reasonably long (from 12 to 20 sentences) - 1 point
	total =	19	

Figure 3: CompQA evaluation framework criteria.

2.2 Question and Argument Collection

We also construct a novel dataset of questions and arguments for our benchmark. First, we randomly select 50 pairs from the Touché dataset (Bondarenko et al., 2022b) and (Chekalina et al., 2021) and retrieve a maximum of 10 arguments for each object from the CAM 2.0 system (Shallouf et al., 2024). More information about these datasets can be found in the Github repository[1]. The exact number varies depending on factors like the popularity of the object, the availability of arguments in its favor on online forums, etc. The pairs from both sources belong to various domains, e.g. companies (*"IBM vs Sony"*), places (*"Virginia vs. Michigan"*), and products of different categories (*"PS3 vs. DS"* and *"tea vs coffee"*), etc. Table 1 provides a few key details about this dataset. On average, each pair contains 10 arguments, with the number ranging from 3 to 20. Only 18 pairs are provided a defined aspect as input. If available, we also extract the questions for the pairs; otherwise, a basic question is used as a default: *"What is better: {object1} or {object2}? Focus on {aspect}."*

Additionally, we have extended the dataset to 432 manually verified questions with extracted objects and aspects using the CompQA dataset of Beloucif et al. (2022). Further benchmark experiments on those datasets are provided in Appendix C.

2.3 Prompt Scenarios for Comparative Answer Generation

We use several prompts with varying levels of complexity and specificity to get summaries from the LLMs for the following reasons:

- check whether our framework can differentiate between good and bad summaries;
- to assess the capability of LLMs to produce comparative answers of good quality with and without provided arguments;
- how the quality of the output answer can vary with prompt-engineering for the same model.

Figure 4 shows the exact content of these scenarios. The **first** scenario is the simplest one that does not include any details from the user other than the objects to compare. It should produce a summary

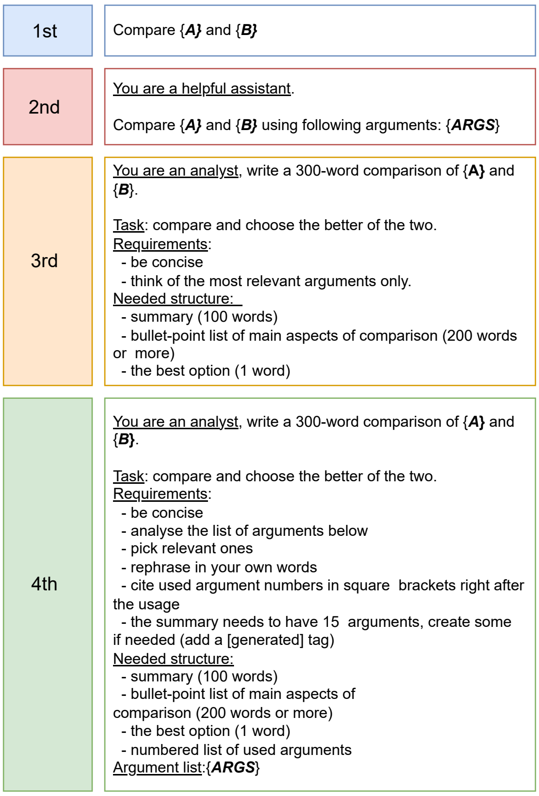


Figure 4: Four prompting scenarios used for comparative summary generation. Objects "A" and "B" are extracted from the input question, "ARGS" is the list of arguments extracted from CAM (Shallouf et al., 2024).

with an arbitrary structure. It is important to note that the remaining three scenarios assign roles to the LLM, whereas the first one does not.

The **second** scenario gives the LLM a list of arguments (ARGS) extracted from CAM, to see if the LLM uses the given arguments exclusively or generates new ones.

The **third** and **fourth** scenarios add more instructions to guide the LLM in producing a summary with a specific structure as described in Section 2. The **third** scenario removes CAM arguments, testing the LLM's ability to generate its own arguments. Lastly, the fourth and most comprehensive scenario includes CAM arguments and improves upon the specific instructions. This configuration is based on the methodology established by Nikishina et al. (2025) to generate ideal comparative summaries. Due to their increased complexity, the third and fourth scenarios aim to produce a summary that should score better than other scenarios.

3 Experimental Setup

In this section, we introduce the models we have selected as potential LLM cores for our agent. Our

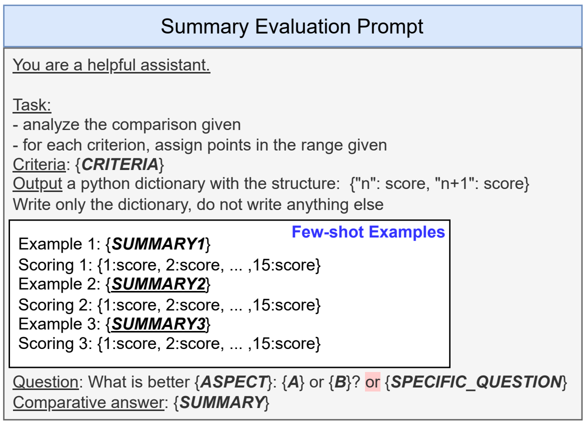


Figure 5: Evaluation prompt for LLMs based on the CompQA evaluation framework criteria. For few-shot, 2 human answers are added to the placeholder.

methodology employs these models both to generate comparative answer datasets using the prompts shown in Figure 4, and to serve as evaluators for our proposed framework. To ensure a rigorous comparison using the CompQA benchmark, we measure their performance alongside manual human assessments.

Participated CQA Agents. Here are short descriptions of LLMs we used for competition in the CompQA Benchmark. We use the standard configuration and parameters for generation:

- **ChatGPT** (GPT-3.5-turbo), details are described by Brown et al. (2020);
- **GPT-4o**, more details in OpenAI (2023);
- **meta-llama/Meta-Llama-3-8B**, more details in Dubey et al. (2024);
- **meta-llama/Meta-Llama-3-70B**, more details in Dubey et al. (2024);
- **mistralai/Mixtral-8x7B-Instruct-v0.1** is a pretrained generative Sparse Mixture of Experts (Jiang et al., 2024);
- **Perplexity AI**[2] — AI-powered research and conversational search engine that answers queries using natural language predictive text, it uses sources from the web and cites links within the text response;

For each LLM, we generate four summaries as per each prompt template shown in 4. We also return three versions of the answer for the GPT family of models (choices, $n = 3$), which are treated and evaluated as separate summaries. In total, our generative pipeline produced **2,000** comparative

[2] https://www.perplexity.ai

answers (comprising 50 examples across 4 scenarios for the 4 models, and 150 examples across 4 scenarios for 2 GPT-based models).

Additional CQA Datasets. In addition to using the LLM agents to generate comparative answers, we also use the following comparative answer datasets in our evaluation benchmark.

- **CAM** dataset comprises the summaries provided by Shallouf et al. (2024). These were produced using the *"lmsys/vicuna-7b-v1"* model for all **50** pairs used in our dataset.
- **Yahoo!Answers** dataset from Chekalina et al. (2021) comprises questions and answers (written by humans). It only contains **28** object pairs from our dataset.
- **Human** dataset – this dataset is sourced from Nikishina et al. (2025) and consists of **80** comparative summaries. These were initially generated by ChatGPT using *Scenario 4* and subsequently refined by four experts in computational linguistics to ensure maximum quality and factual accuracy.

Obtaining Automatic Assessments using LLMs. We automatically evaluated 2,158[3] answers with all participating LLMs based on the CompQA evaluation framework using the 2-shot prompting.

Obtaining Assessments from Human Annotators. To verify the quality of automatic annotation, a total of *367* answers are randomly selected and manually reviewed by expert annotators (the authors). *123* of them are done with an overlap of two annotators to measure the annotation agreement. Krippendorf's alpha is equal to **0.75** for the final score and **0.71** for all scores on average.

3.1 Comparison of LLM and Human Assessments

In order to understand whether the human annotation can be replaced with an automatic one, we compare the agreement and annotation scores between different LLMs and human answers. Table 2 demonstrates both the agreement score (Krippendorff's alpha) and Spearman correlation scores, following Bavaresco et al. (2024). The results give a positive answer to the **RQ-2**: "LLMs can reliably evaluate comparative summaries with human expert-level quality."

[3] $4scenarios \cdot 150_{GPT-3.5} + 4scenarios \cdot 150_{GPT-4} + 28_{Yahoo} + 50_{CAM} + 80_{Human} + 4scenarios \cdot 50 \cdot 4_{Llama3-8B,Llama3-70B,Mixtral,Perplexity} = 2158$

Model	α	**Spearman's**
GPT-4o, separately	**0.71**	0.69, $p < 0.001$
GPT-4o, final score	**0.58**	0.55, $p < 0.001$
GPT-3.5, separately	0.63	0.63, $p < 0.001$
GPT-3.5, final score	0.31	0.40, $p < 0.002$
Perplexity, separately	**0.71**	0.72, $p < 0.001$
Perplexity, final score	0.39	0.60, $p < 0.001$
Mixtral, separately	0.55	0.69, $p < 0.001$
Mixtral, final score	0.22	0.62, $p < 0.001$
Llama-3-8B, separately	0.48	0.49, $p < 0.001$
Llama-3-8B, final score	0.41	0.46, $p < 0.001$
Llama-3-70B, separately	0.61	**0.76**, $p < 0.001$
Llama-3-70B, final score	0.47	**0.72**, $p < 0.001$

Table 2: Agreement (Krippendorf's α) and correlation scores between human and LLM evaluations. Separate scores are calculated for all scores concatenated for all answers (human annotation against model annotation), the total scores represent the sums of 15 criteria (denoted in Figure 3) for each answer.

From the results in Table 2, we can see that both agreement and correlation scores are much higher for the separate scores than for the summary scores. Moreover, we can also conclude that Llama-3-70B is better according to Spearman's correlation, while GPT-4o and Perplexity show higher agreement according to Krippendorff's α.

4 CQA Systems Performance

In this section, we present the results of existing CQA agents on our benchmark, and answer the **RQ3**: "How do different LLMs fare against each other in generating high-quality summaries?" Table 3 shows the best results obtained by each comparative answer dataset (row) on our benchmark with different evaluators (column). For human evaluations, scores are reported for a randomly sampled dataset of 367 summaries. The results of all LLM evaluators unanimously rate the GPT-4o answers as the first. The second place is shared between Mixtral and Perplexity. Regarding human evaluations, it is evident that the highest quality responses are those generated by humans (**16.69** on average), while the best LLM answers are still generated by GPT-4. Interestingly, the answers ranked as the lowest belong to Yahoo and CAM datasets, primarily because these datasets do not align with the structural criteria used by our evaluation framework. To see the detailed results, refer to Table 5 in Appendix A.

	GPT-4o	GPT-3.5	Perplexity	Mixtral	Llama-3-8B	Llama-3-70B	Human*
GPT-4o	**17.77**$_{\pm 1.06}$	**17.54**$_{\pm 2.14}$	**18.69**$_{\pm 0.69}$	**18.45**$_{\pm 2.64}$	**16.64**$_{\pm 1.86}$	**18.40**$_{\pm 0.70}$	15.96$_{\pm 1.35}$
GPT-3.5	16.66$_{\pm 1.86}$	16.05$_{\pm 2.62}$	16.12$_{\pm 2.64}$	18.05$_{\pm 1.33}$	15.08$_{\pm 3.18}$	16.25$_{\pm 2.32}$	14.58$_{\pm 2.34}$
Perplexity	17.42$_{\pm 1.34}$	16.52$_{\pm 3.00}$	18.47$_{\pm 1.11}$	18.34$_{\pm 0.87}$	16.31$_{\pm 2.06}$	17.87$_{\pm 1.15}$	15.00$_{\pm 1.77}$
Mixtral	17.20$_{\pm 1.31}$	17.05$_{\pm 2.43}$	17.92$_{\pm 1.41}$	18.36$_{\pm 0.94}$	16.29$_{\pm 2.40}$	17.45$_{\pm 1.67}$	13.75$_{\pm 2.63}$
Llama-3-8B	16.58$_{\pm 1.80}$	15.95$_{\pm 2.60}$	17.19$_{\pm 1.81}$	18.08$_{\pm 1.24}$	15.56$_{\pm 2.72}$	16.15$_{\pm 1.90}$	14.77$_{\pm 2.16}$
Llama-3-70B	17.12$_{\pm 1.37}$	16.22$_{\pm 2.76}$	17.93$_{\pm 1.39}$	18.19$_{\pm 1.068}$	15.70$_{\pm 2.95}$	17.27$_{\pm 1.45}$	15.37$_{\pm 1.72}$
CAM (Shallouf et al., 2024)	9.52$_{\pm 3.808}$	8.74$_{\pm 4.56}$	6.46$_{\pm 2.89}$	9.66$_{\pm 4.43}$	6.54$_{\pm 4.21}$	6.40$_{\pm 2.29}$	8.04$_{\pm 2.78}$
Yahoo (Chekalina et al., 2021)	9.32$_{\pm 3.95}$	5.88$_{\pm 4.86}$	5.96$_{\pm 3.68}$	13.32$_{\pm 4.97}$	5.21$_{\pm 5.91}$	5.96$_{\pm 3.64}$	6.00$_{\pm 1.00}$
Human	17.09$_{\pm 1.72}$	15.29$_{\pm 2.68}$	18.33$_{\pm 1.03}$	17.65$_{\pm 1.57}$	14.29$_{\pm 3.23}$	17.96$_{\pm 1.56}$	**16.69**$_{\pm 1.76}$

Table 3: Average scores for all participating models for LLM and human evaluations. Rows are the datasets. Columns represent the evaluation models. Human* evaluations were conducted on a random subset.

Comparison between scenarios. Figure 6 shows average scores for all models. Scenarios 3 and 4 achieved the highest scores from both human and Llama-3-70B evaluators. Conversely, the summaries generated by Scenarios 1 and 2 are consistently ranked lowest. This highlights the fact that the evaluation framework can help to differentiate between summaries of varied structures. The fourth scenario uses arguments from CAM, and they have scored lower than their similar counterpart without these arguments i.e. the third scenario, which is more pronounced in the human evaluation. This discrepancy raises concerns about the quality of CAM arguments, despite the prompt explicitly requiring the selection of only relevant arguments. Thus, we further compare these models for all criteria of our evaluation framework, by grouping them into *structure*, *quality*, and *relevance* in Figure 7. The comparison shows that the summaries generated with the third and fourth scenarios scored higher for all categories of our framework. The second scenario suffered noticeably for *structure*, this difference is less noticeable between the third and fourth scenarios.

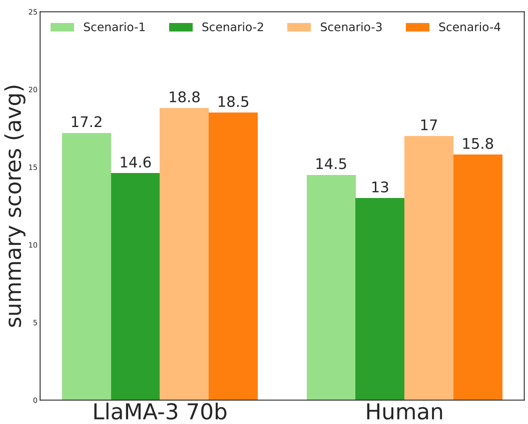


Figure 6: Average scores for each scenario for Llama-3-70B and manual human assessments.

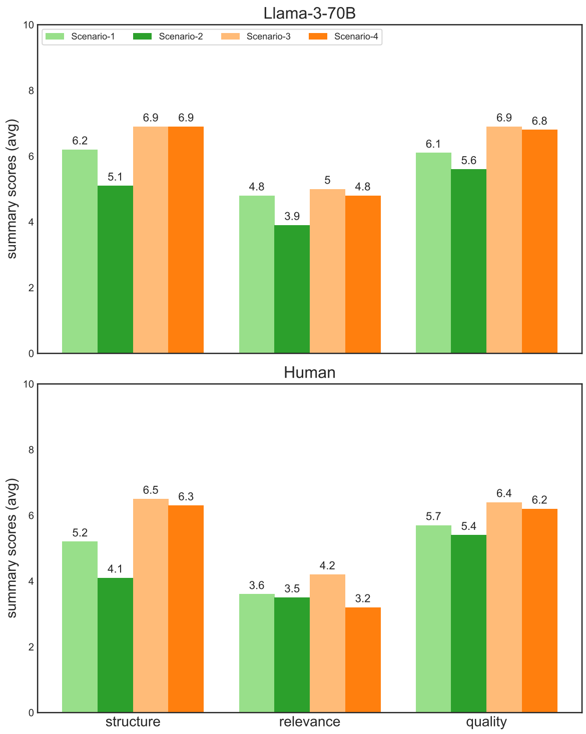


Figure 7: Average scores for each scenario, distributed into three categories of evaluation framework: **structure**–(7 points), **quality**–(5 points), and **relevance**–(7 points).

Comparison between Llama-3-70B and human: Human evaluations are consistently lower than Llama-3-70B. Table 4 shows that Llama-3-70B performs similarly for the questions with and without aspect, while the human scores are lower when the aspect is presented, which implies that Llama-3-70B might not give the importance to the required aspect when evaluating the summary.

Additional analysis is also present in Figure 8: we aim to check, whether the scores assigned by Llama-3-70B to the answers evaluated by humans for the extreme cases are coherent. Human evaluation shows average results (with means at 14-17) for the cases evaluated as good ones by Llama-3-70B, and also low results for the answers assigned

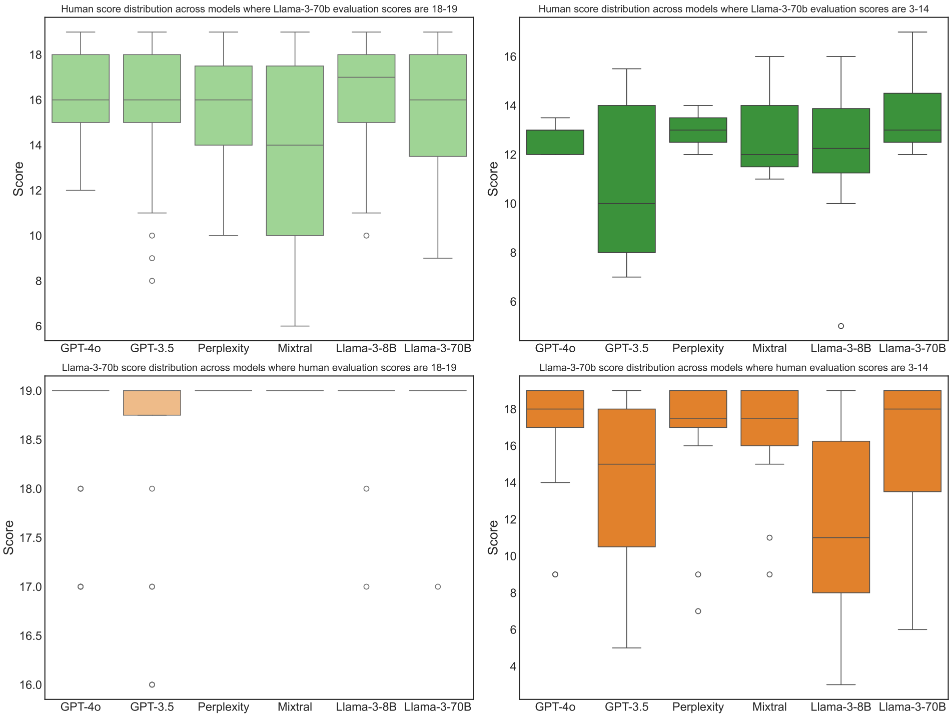


Figure 8: Distribution scores across models for the highest and lowest scores assigned by Llama-3-70B and Human evaluations.

Aspect	Llama-3-70B	Human
yes	$16.43_{\pm 3.52}$	$13.43_{\pm 2.98}$
no	$16.44_{\pm 3.52}$	$15.01_{\pm 2.69}$

Table 4: Average scores of human and Llama-3-70B, for the subset of answers **with** and **without** aspect.

with low scores by the LLM. It means that the model slightly tends to overestimate the performance of the LLMs in comparison to the human evaluation. From the second row, we can see that Llama-3-70B assigns high scores for the answers which are also ranked high by humans (lower left subfigure) and both high and low scores for the answers ranked as "bad" (3-14) by humans (lower right subfigure).

Distilling a Better Open-Source CQA Agent Our initial results indicate that the highest performance on the benchmark is currently dominated by proprietary, closed-source models such as GPT-4o and Perplexity (Table 3). In contrast, Llama-3-8B-Instruct, the smallest LLM used, performs decently. Motivated by its potential, we investigate whether supervised fine-tuning (SFT) with LoRA adapter can close the gap between this open-source baseline and top closed-source models. We compiled the training data using the answers from the models gaining $>=$ 18 points. The training data comprises 1244 question-answer pairs in total. The best SFT variant scored $\mathbf{18.40}_{\pm 1.25}$, which performs on par with GPT-4o; detailed fine-tuning parameters are provided in Appendix B. Overall, the results show that distilling the capabilities of larger models into smaller ones remains promising.

Additional Experiments We have also conducted additional experiments on reasoning LLMs, extended dataset, and possible biases that are presented in Appendix C.

5 Related Work

In this section, we briefly introduce each subtask for Comparative Question Answering and also discuss the existing LLM evaluation benchmarks.

5.1 Comparative Question Answering

Here, we introduce each subtask and list several papers that addressed the topic.

Comparative Question Identification aims at classifying questions into two types: comparative and non-comparative. This classification task is solved with both Encoder and Decoder Transformer models (Bondarenko et al., 2020, 2022a; Shallouf et al., 2024). **Object and Aspect Identification** is a sequence labelling task, aims at finding objects and aspect of comparison in the question. There exist varios datasets and approaches to solve the task, mostly, with Trasnformer models (Chekalina et al., 2021; Beloucif et al., 2022; Bondarenko et al., 2022a; Shallouf et al., 2024). **Stance Classification** is another classification task, that identifies the stance of comparative sentences. Panchenko et al. (2019), Bondarenko et al. (2022a), and Kang et al. (2023) solve the task using standard ML classifier, Encoder-based Transformer, and GPT-4o respectively. **Summary Generation** is only partially tackled by Chekalina et al. (2021) and Shallouf et al. (2024). The closest work on multi-document summarization of differing opinions is by Iso et al. (2022), which focuses on aggregating diverse opinions and synthesizing them into a coherent summary.

5.2 LLM Evaluation Benchmarks

Apart from the well-known benchmarks like SuperGLUE (Sarlin et al., 2020), MTEB (Muennighoff et al., 2023), and SQuAD (Rajpurkar et al., 2016), several more challenging benchmarks have gained prominence: MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), BIG-Bench Hard (Suzgun et al., 2023). The LLM evaluation framework proposed by (Chiang and Lee, 2023) involves presenting a Large Language Model with task instructions, a sample to be evaluated and a question. The researchers use LLM evaluation to score parts of both generated and human-written stories. To facilitate comprehensive evaluations, several initiatives aggregate multiple benchmarks: Hugging Face's Big Benchmarks Collection (a centralized platform for various leaderboards), and LMSys Chatbot Arena (Chiang et al., 2024; Zheng et al., 2023): it also utilizes user ratings and GPT-4o evaluations to assess chatbot performance.

6 Conclusion

We have proposed a comprehensive CompQA evaluation framework based on 15 various criteria to be used in scoring comparative answers, demonstrating that GPT-4 produces the best answers to comparative questions. We have manually evaluated the answers from a range of models and datasets using our framework and compared human scores with LLM evaluations, showing that LLM results have a strong correlation with human expert-level evaluations. Moreover, we have trained a smaller model Llama-3-8B-Instruct to generate summaries at the similar level of quality as closed models.

7 Limitations

The main limitations of the paper are as follows:

- Regarding the human evaluation dataset, we acknowledge that our sample size is quite small and might be extended even further to make comparisons of better quality.
- The arguments used in the strategies are solely derived from the CAM framework. This reliance on a single source may have constrained the retrieval performance, particularly in Scenario 2 and 4, leading to suboptimal results. To address this, future work could involve annotators in crafting or refining arguments, which may enhance the robustness and effectiveness of the strategies.

8 Ethical Considerations

In our benchmark we test multiple LLMs, one key concern is the handling of user data by proprietary models like OpenAI's GPT-4, which are developed and maintained by private companies. These companies often retain the right to use input data for improving their models, as stated in their terms of service and privacy policies. As a result, personal or sensitive information provided by users during interactions with these models could be logged, stored, or analyzed for commercial purposes. While companies may anonymize data, the potential use of personal information in ways that users may not fully understand or consent to raises significant privacy concerns.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *CoRR*, abs/2406.18403.

Meriem Beloucif, Seid Muhie Yimam, Steffen Stahlhacke, and Chris Biemann. 2022. Elvis vs. M. Jackson: Who has more albums? classification and identification of elements in comparative questions. In *Proceedings of the Thirteenth Language Resources* *and Evaluation Conference*, pages 3771–3779, Marseille, France. European Language Resources Association.

Alexander Bondarenko, Yamen Ajjour, Valentin Dittmar, Niklas Homann, Pavel Braslavski, and Matthias Hagen. 2022a. Towards understanding and answering comparative questions. In *WSDM*, pages 66–74. ACM.

Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. 2020. Comparative web search questions. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 52–60.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022b. Overview of Touché 2022: Argument Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association (CLEF 2022)*, volume 13390 of *Lecture Notes in Computer Science*, Berlin Heidelberg New York. Springer.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Viktoriia Chekalina, Alexander Bondarenko, Chris Biemann, Meriem Beloucif, Varvara Logacheva, and Alexander Panchenko. 2021. Which is better for deep learning: Python or MATLAB? answering comparative questions in natural language. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *Preprint*, arXiv:2304.02554.

Som Gupta and S. K Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.

Inwon Kang, Sikai Ruan, Tyler Ho, Jui-Chien Lin, Farhad Mohsin, Oshani Seneviratne, and Lirong Xia. 2023. Llm-augmented preference learning from natural language. *CoRR*, abs/2310.08523.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

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.

Manikanta Loya, Divya Sinha, and Richard Futrell. 2023. Exploring the sensitivity of LLMs' decision-making capabilities: Insights from prompt variations and hyperparameters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3711–3716, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Irina Nikishina, Saba Anwar, Nikolay Dolgov, Maria Manina, Daria Ignatenko, Artem Shelmanov, and Chris Biemann. 2025. How to compare things properly? a study of argument relevance in comparative question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15702–15720, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing comparative sentences. In *Proceedings of the 6th Workshop on Argument Mining*, pages 136–145, Florence, Italy. Association for Computational Linguistics.

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. Superglue: Learning feature matching with graph neural networks. *Preprint*, arXiv:1911.11763.

Hassan Shakil, Ahmad Farooq, and Jugal Kalita. 2024. Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, 603:128255.

Ahmad Shallouf, Hanna Herasimchyk, Mikhail Salnikov, Rudy Alexandro Garrido Veliz, Natia Mestvirishvili, Alexander Panchenko, Chris Biemann, and Irina Nikishina. 2024. CAM 2.0: End-to-end open domain comparative question answering system. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2657–2672, Torino, Italia. ELRA and ICCL.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Complete Results

Table 5 shows the performance of all models, for each scenario.

Model	Scenario	Mixtral	GPT-4o	Human	Llama-3-8B	Llama-3-70B	Perplexity	GPT-3.5
GPT-3.5	1	18.76$_{\pm 0.59}$	16.95$_{\pm 1.44}$	14.93$_{\pm 2.84}$	17.07$_{\pm 2.14}$	16.63$_{\pm 1.27}$	16.84$_{\pm 2.14}$	17.90$_{\pm 1.65}$
GPT-3.5	2	16.77$_{\pm 2.80}$	14.10$_{\pm 3.15}$	11.46$_{\pm 2.93}$	11.86$_{\pm 4.00}$	12.14$_{\pm 4.99}$	11.07$_{\pm 5.79}$	13.39$_{\pm 4.07}$
GPT-3.5	**3**	18.71$_{\pm 0.57}$	18.21$_{\pm 1.02}$	16.92$_{\pm 1.90}$	16.59$_{\pm 3.09}$	18.54$_{\pm 1.02}$	18.44$_{\pm 1.12}$	17.74$_{\pm 1.56}$
GPT-3.5	4	18.02$_{\pm 1.35}$	17.37$_{\pm 1.83}$	15.00$_{\pm 1.70}$	14.81$_{\pm 3.51}$	17.69$_{\pm 2.00}$	18.12$_{\pm 1.53}$	15.17$_{\pm 3.21}$
GPT-4o	1	18.57$_{\pm 0.72}$	17.59$_{\pm 1.00}$	15.22$_{\pm 1.52}$	17.44$_{\pm 1.53}$	17.90$_{\pm 0.86}$	18.41$_{\pm 0.88}$	17.84$_{\pm 2.15}$
GPT-4o	2	17.97$_{\pm 8.75}$	17.22$_{\pm 1.33}$	14.48$_{\pm 1.62}$	14.56$_{\pm 2.78}$	17.75$_{\pm 1.55}$	18.41$_{\pm 1.50}$	16.77$_{\pm 2.76}$
GPT-4o	**3**	18.91$_{\pm 0.31}$	18.23$_{\pm 0.95}$	17.24$_{\pm 1.33}$	17.71$_{\pm 1.53}$	18.95$_{\pm 0.32}$	18.94$_{\pm 0.31}$	18.36$_{\pm 1.14}$
GPT-4o	4	18.38$_{\pm 0.78}$	18.03$_{\pm 0.95}$	16.89$_{\pm 0.94}$	16.83$_{\pm 1.59}$	18.99$_{\pm 0.08}$	18.99$_{\pm 0.08}$	17.21$_{\pm 2.47}$
Llama-3-70B	1	18.74$_{\pm 0.56}$	17.22$_{\pm 1.20}$	14.04$_{\pm 1.59}$	17.34$_{\pm 2.16}$	17.34$_{\pm 1.00}$	18.06$_{\pm 1.00}$	17.38$_{\pm 2.66}$
Llama-3-70B	2	16.44$_{\pm 2.62}$	15.90$_{\pm 1.78}$	13.29$_{\pm 2.40}$	12.70$_{\pm 4.11}$	13.86$_{\pm 4.22}$	15.74$_{\pm 4.30}$	13.69$_{\pm 4.11}$
Llama-3-70B	**3**	18.76$_{\pm 0.59}$	18.06$_{\pm 1.11}$	17.69$_{\pm 1.58}$	16.82$_{\pm 2.67}$	18.96$_{\pm 0.20}$	19.00$_{\pm 0.00}$	17.90$_{\pm 1.31}$
Llama-3-70B	4	18.80$_{\pm 0.49}$	17.30$_{\pm 1.37}$	16.46$_{\pm 1.30}$	15.96$_{\pm 2.86}$	18.90$_{\pm 0.36}$	18.92$_{\pm 0.27}$	15.94$_{\pm 2.95}$
Llama-3-8B	1	18.70$_{\pm 0.71}$	16.78$_{\pm 1.40}$	14.00$_{\pm 2.94}$	16.31$_{\pm 2.36}$	16.52$_{\pm 1.71}$	17.56$_{\pm 0.93}$	17.67$_{\pm 2.15}$
Llama-3-8B	2	16.12$_{\pm 3.15}$	14.10$_{\pm 3.36}$	11.94$_{\pm 2.65}$	11.51$_{\pm 4.90}$	11.04$_{\pm 4.40}$	13.52$_{\pm 5.41}$	11.76$_{\pm 4.83}$
Llama-3-8B	**3**	18.92$_{\pm 0.27}$	18.26$_{\pm 0.88}$	16.50$_{\pm 1.96}$	18.60$_{\pm 0.64}$	19.00$_{\pm 0.00}$	19.00$_{\pm 0.000}$	18.58$_{\pm 0.70}$
Llama-3-8B	4	18.59$_{\pm 0.81}$	17.18$_{\pm 1.57}$	16.68$_{\pm 1.10}$	15.82$_{\pm 2.98}$	18.02$_{\pm 1.49}$	18.68$_{\pm 0.89}$	15.81$_{\pm 2.73}$
Mixtral	1	18.71$_{\pm 0.62}$	16.80$_{\pm 1.20}$	14.67$_{\pm 1.73}$	17.16$_{\pm 1.81}$	16.58$_{\pm 2.11}$	17.48$_{\pm 0.79}$	17.31$_{\pm 3.05}$
Mixtral	2	17.50$_{\pm 1.98}$	16.12$_{\pm 1.76}$	11.54$_{\pm 3.66}$	13.74$_{\pm 3.81}$	15.54$_{\pm 3.48}$	16.54$_{\pm 3.78}$	15.73$_{\pm 3.32}$
Mixtral	**3**	18.86$_{\pm 0.50}$	18.40$_{\pm 0.86}$	16.29$_{\pm 2.73}$	18.02$_{\pm 1.86}$	18.86$_{\pm 0.53}$	18.84$_{\pm 0.51}$	18.06$_{\pm 1.39}$
Mixtral	4	18.38$_{\pm 0.70}$	17.48$_{\pm 1.43}$	12.50$_{\pm 2.39}$	16.22$_{\pm 2.13}$	18.82$_{\pm 0.56}$	18.80$_{\pm 0.57}$	17.12$_{\pm 1.97}$
Perplexity	1	18.78$_{\pm 0.46}$	17.32$_{\pm 1.24}$	12.86$_{\pm 1.48}$	17.28$_{\pm 1.98}$	17.68$_{\pm 0.91}$	18.08$_{\pm 0.99}$	17.22$_{\pm 3.75}$
Perplexity	2	16.84$_{\pm 2.33}$	16.46$_{\pm 2.07}$	14.14$_{\pm 2.00}$	13.88$_{\pm 2.97}$	16.00$_{\pm 2.79}$	18.04$_{\pm 2.18}$	14.52$_{\pm 3.78}$
Perplexity	**3**	19.00$_{\pm 0.00}$	18.68$_{\pm 0.47}$	17.40$_{\pm 1.84}$	18.40$_{\pm 1.03}$	19.00$_{\pm 0.00}$	19.00$_{\pm 0.00}$	18.53$_{\pm 0.96}$
Perplexity	4	18.76$_{\pm 0.69}$	17.20$_{\pm 1.56}$	15.64$_{\pm 1.75}$	15.68$_{\pm 2.27}$	18.78$_{\pm 0.91}$	18.76$_{\pm 1.29}$	15.82$_{\pm 3.53}$
Human		17.65$_{\pm 1.58}$	17.088$_{\pm 1.72}$	16.69$_{\pm 1.76}$	14.29$_{\pm 3.23}$	17.96$_{\pm 1.56}$	18.33$_{\pm 1.03}$	15.29$_{\pm 2.68}$
Yahoo Chekalina et al. (2021)	-	13.32$_{\pm 4.97}$	9.32$_{\pm 3.95}$	6.00$_{\pm 1.00}$	5.21$_{\pm 5.91}$	5.96$_{\pm 3.64}$	5.96$_{\pm 3.68}$	5.88$_{\pm 4.86}$
CAM Shallouf et al. (2024)		9.66$_{\pm 4.43}$	9.52$_{\pm 3.81}$	8.04$_{\pm 2.78}$	6.54$_{\pm 4.21}$	6.40$_{\pm 2.29}$	6.46$_{\pm 2.89}$	8.74$_{\pm 4.56}$

Table 5: CompQA Benchmark leaderboard for all participating models, against each scenario.

B Training Details

In this section, we present the details of supervised fine-tuning. It was performed on 4 NVIDIA A100 GPUs using DeepSpeed with ZeRO Stage 3 (Rajbhandari et al., 2020) and the Adam optimizer. We used a learning rate of 1e-6, warmup ratio 0.03, 3 epochs, a batch size of 32, weight decay 0.01, with a cosine annealing scheduler.

C Additional Experiments

Reasoning LM Performance Additionally, we generated the comparative answers for the following reasoning language models: o3-mini, DeepSeek-R1, Qwen3-8B (with thinking enabled), and Gemini-2.5-Flash. To provide a comprehensive comparison with traditional models, we also include Gemma-3-27B-it, Qwen3-8B (with thinking disabled), along with GPT-4o (drawn from Table 3). For the evaluation, we use the Llama-3-70B model, which is most correlated with human responses. From the results in Table 6, we can see that DeepSeek-R1 achieves the best score, while Qwen3-8B does not perform well in both reasoning and non-reasoning options. In general, we can also see that models with reasoning enabled perform better. Importantly, these findings support the main contributions of this paper: the developed criteria are extensible to newer models.

Results on the Extended Dataset To check, whether the scores for the main benchmark dataset are consistent with the extended one, we have evaluated the best performing LLMs on the 432 questions carefully filtered from the dataset of Beloucif et al. (2022). It is important to note that 173 questions do not have any arguments retrieved from the CAM system; therefore, the scores for the second and fourth scenarios are noisy. From Table 7, it can be seen that for the non-reasoning models, GPT4-o yields the best results, while for the reasoning models, both DeepSeek-R1 and o3-mini perform the best.

model	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Total score
GPT-4o (from Table 3)	$17.90_{\pm 0.86}$	$17.75_{\pm 1.55}$	$18.95_{\pm 0.32}$	$18.99_{\pm 0.08}$	$18.40_{\pm 0.70}$
Gemma3-27B-it	$16.62_{\pm 1.53}$	$14.32_{\pm 5.48}$	$16.96_{\pm 0.98}$	$18.48_{\pm 1.04}$	$16.60_{\pm 3.29}$
Qwen3-8B (disabled thinking)	$14.12_{\pm 4.96}$	$11.02_{\pm 6.74}$	$17.00_{\pm 2.16}$	$16.08_{\pm 5.80}$	$14.68_{\pm 5.74}$
o3-mini	$17.36_{\pm 1.21}$	$17.80_{\pm 1.22}$	$18.96_{\pm 0.28}$	$19.00_{\pm 0.00}$	$18.28_{\pm 1.13}$
DeepSeek-R1	$18.20_{\pm 0.92}$	$18.24_{\pm 1.09}$	$18.94_{\pm 0.31}$	$19.00_{\pm 0.00}$	$\mathbf{18.60}_{\pm 0.81}$
Gemini-2.5-Flash	$18.16_{\pm 1.05}$	$17.14_{\pm 3.19}$	$18.92_{\pm 0.34}$	$18.92_{\pm 0.39}$	$18.29_{\pm 1.85}$
Qwen3-8B (enabled thinking)	$14.98_{\pm 3.75}$	$13.52_{\pm 6.33}$	$17.58_{\pm 3.23}$	$16.08_{\pm 5.96}$	$15.54_{\pm 5.22}$

Table 6: The results on comparison between reasoning LLMs and traditional autoregressive LLMs using the Llama-3-70B model for evaluation.

model	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Total score
Llama-3-70B	$18.88_{\pm 0.90}$	$18.17_{\pm 2.53}$	$18.95_{\pm 0.32}$	$18.88_{\pm 0.58}$	$18.72_{\pm 1.42}$
GPT-4o	$18.98_{\pm 0.21}$	$18.90_{\pm 1.02}$	$18.97_{\pm 0.27}$	$18.92_{\pm 0.75}$	$18.94_{\pm 0.66}$
o3-mini	$18.92_{\pm 1.03}$	$18.81_{\pm 1.36}$	$19.00_{\pm 0.00}$	$18.91_{\pm 0.29}$	$18.91_{\pm 0.87}$
DeepSeek-R1	$18.92_{\pm 0.92}$	$18.81_{\pm 1.27}$	$18.97_{\pm 0.26}$	$18.92_{\pm 0.28}$	$18.91_{\pm 0.81}$
Llama-3-8B LoRA finetuning	-	-	-	$18.59_{\pm 2.10}$	-

Table 7: The results on the extended dataset using the Llama-3-70B model for evaluation.

Possible Biases Inherent to Certain LLMs. Another possible direction that should be worked on to improve upon this research paper is finding any biases that could exist within some LLMs.

An example of this that we encountered while generating summaries was that GPT-4, when asked to compare Google and Yahoo search engines and given arguments favoring the latter, created comparisons strongly for Google (see examples on Google vs. Yahoo in the Github repository. Moreover, the model rewrote the arguments, hallucinating that they were all originally favoring Google. We argue that such cases of bias where the final answer is not derived from the listed arguments could be because of the model inner bias towards a more popular object or subject.

Needless to say, this kind of favoritism embedded in an LLM may heavily impact its objectivity, and further tests need to be conducted to determine the cause and possible solutions to this issue. This is an exception, however, as no other object pair created such a situation; this means that, hopefully, this clear bias is something exceedingly rare.

Illustrating Arguments with Images Using Aspect-Aware Prompting

Maximilian Heinrich
Bauhaus-Universität Weimar
Weimar, Germany
maximilian.heinrich@uni-weimar.de

Sharat Anand
Bauhaus-Universität Weimar
Weimar, Germany
sharat.anand@uni-weimar.de

Johannes Kiesel
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany
johannes.kiesel@gesis.org

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany
benno.stein@uni-weimar.de

Abstract

Images can powerfully strengthen arguments, conveying ideas more immediately and compellingly than text alone. With the rise of text-to-image models, a broad audience can now generate custom visuals to illustrate their arguments. Yet a fundamental mismatch undermines this potential: these models are trained on concrete scene descriptions, while arguments operate at the level of general, abstract principles. Naively prompting such a model with an argumentative text therefore rarely produces images that genuinely illustrate the argument. To address this challenge, we propose an aspect-aware image generation approach. Given an argument, our method first identifies the key aspects that an illustrative image should convey, then constructs a detailed scene description grounded in both the argument and those aspects, and finally generates an image using that scene description as the prompt. A human-assessment evaluation demonstrates that this approach yields images that illustrate arguments significantly better than those produced by naive prompting.

1 Introduction

Images are ubiquitous in contemporary life. From social media and news feeds to advertising billboards and political campaigns, images have become a prominent mode of communication and argumentation (Groarke, 2024). Images can help to understand and retain complex information (Clark and Paivio, 1991). Moreover, images can evoke strong emotional responses—both positive and negative—and suggest or motivate particular actions (Freedberg, 1989; Dunaway, 2018). In advertising, political campaigns, and public discourse, images are frequently used to support judgments, justify positions, and motivate action. In this sense, they serve as visual arguments that can be as persuasive as their verbal counterparts (Groarke, 1996). However, until recently, most people could not illustrate their arguments with specially tailored images due to the high costs of generating images.[1]

Despite recent advances in text-to-image generation, the generation of images that effectively convey arguments remains a challenge. Whereas arguments are typically expressed at an abstract and conceptual level, text-to-image models are trained on concrete descriptions of scenes involving identifiable objects, actions, and settings. As a result of this mismatch, images generated with text-to-image models using an argument as a literal prompt are incomplete, inconsistent, or unrelated to the intended claim. The columns "Argument as literal prompt" in Table 1 (on the right-hand side) exemplify this problem: key aspects of the arguments are missing in the respective images.

To address this challenge, we propose aspect-aware prompt generation for generating images that illustrate arguments. The proposed approach comprises three steps from an argument to the image: (1) identify the argument's key aspects—visually interpretable words and phrases that capture the key ideas of an argument; (2) generate a descriptive scene-level prompt from the argument and the identified aspects; and (3) generate the image using the generated prompt. Column "Aspect-aware

[1] Internet memes are somewhat of an exception, as they – to put it simply – allow a generic visual pattern to be adapted to a specific argument. But memes cannot cover all arguments.

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 72–85
July 3, 2026 ©2026 Association for Computational Linguistics

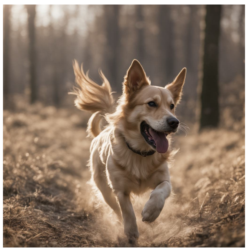
Argument	Aspect-aware prompt	Argument as literal prompt	
Sugar makes food more enjoyable.			
Dogs require regular outdoor activity which promotes a healthier lifestyle for their owners.			
Risk of crime is higher in public transport.			
	Stable Diffusion 1.0 XL	Stable Diffusion 1.0 XL	Stable Diffusion 3.5 Medium

Table 1: Turning an argument into an image: Using the literal argument (first column) as an image prompt results in images that do not reflect the essence of the argument (third and fourth columns), while the second column shows the generated images for which the prompt was created using our aspect-aware approach to prompt generation. Table 2 (third column) shows the prompts that were generated. The image generation models are Stable Diffusion 1.0 XL (second and third columns) and Stable Diffusion 3.5 Medium (fourth column).

prompt" in Table 1 exemplifies the images created for the different arguments mentioned above, showing a clear improvement over the images generated with naive prompting.[2]

This paper is structured as follows: Section 3 discusses the task and challenges of image generation to illustrate arguments in detail. Section 4 presents a new dataset of arguments and associated images for evaluating image generation algorithms. Section 5 introduces seven automatic quality indicators capturing different properties of prompts and generated images—such as imageability—and uses them alongside human assessment to compare images generated from literal prompts against those generated with aspect-aware prompts. The section further evaluates whether the automatically identified aspects align with those identified by humans. Section 6 then discusses the results of the experiments. Our experiments show that prompts derived from argument aspects produce images that convey the argument more effectively than prompts generated directly from the literal argument text.

2 Related Work

The persuasive and cognitive impact of visuals has been documented across numerous domains. A historical overview of the influence of images on human emotions and societal practices is provided by Freedberg (1989). For comprehensive surveys of text-to-image generation methods, we refer the reader to Alhabeeb and Al-Shargabi (2024); Bie et al. (2025). The capability of generative models to respond to and interpret different prompts has been systematically investigated by Liu and Chilton (2022), who analyze how prompt design influences the quality and characteristics of generated images.

Beyond standard generation tasks, image generation has also been employed as an intermediate reasoning step for language models solving complex inferential problems (Chern et al., 2025). Related

[2] Code and data available at: https://github.com/webis-de/ArgMining-26

work further investigates to what extent models can identify visual elements that are relevant to an argument. For instance, Chung et al. (2024) introduce a benchmark for visual argument understanding that reveals selective vision—identifying which visual elements are relevant to an argument—as a key challenge for current models.

At the intersection of argumentation and visualization, one line of research focuses on detecting persuasive techniques in multimodal content such as memes (Dimitrov et al., 2021). Another thread concerns the visual representation of argumentative structures: Khartabil et al. (2021) design visualization techniques for exploring individual argument structures, while Kiesel (2025) develops visual analytics approaches for argumentation at the corpus level. A related direction addresses the generation of complex statistical infographics from unstructured text, requiring models to plan chart types, layouts, and visual structure from textual content alone (Ghosh et al., 2025).

The most direct precedent for our work is the Touché Shared Task on Argument Image Retrieval and Generation (Kiesel et al., 2025, 2024; Bondarenko et al., 2023, 2022), which asks participants to retrieve or generate images that effectively convey a given argument. Existing approaches focus almost exclusively on retrieval from a predefined image pool, with generative models playing only an auxiliary role — for instance, to support ranking via comparison with generated images (Moebius et al., 2023) or to produce synthetic training data for retrieval models (Janusko et al., 2024). We participated in the most recent edition of this shared task with the aspect-aware generation approach presented in this paper (Kiesel et al., 2025; Anand and Heinrich, 2025). However, our shared-task contribution is limited to a system description and does not include a systematic evaluation of the approach — a gap that the present paper addresses.

3 From Arguments to Image Prompts

This section discusses the conceptual foundations of generating images for arguments, first examining the relationship between images and arguments, and then contrasting abstract argumentative statements with concrete scene descriptions.

Images and Arguments It is a fundamental characteristic of images that they are ambiguous yet rich in information (Kjeldsen, 2015). Consider a person standing on a ladder. An image of this situation conveys far more details than a verbal description, such as the person's appearance, the surrounding environment, or the color and material of the ladder. While visual richness provides extensive information, it also introduces ambiguity: it may be unclear whether the person is climbing up or down the ladder, leaving the situation open to interpretation. Another challenge concerns the interpretation of symbolic elements in images. Understanding such symbols often depends on cultural and contextual knowledge that is not explicitly represented in the image itself. Because images are rich in detail yet ambiguous, their meaning is often guided by accompanying textual elements that help anchor the viewer's interpretation and direct attention toward the intended message (Kjeldsen, 2015). The ambiguity of images raises the question of whether visuals can function as arguments at all. Critics argue that identifying argument structures in visuals—such as premise-conclusion relations—depends on background knowledge and interpretive processes that are fundamentally verbal in nature (Johnson, 2003). Others argue that similar ambiguity is also present in textual arguments (Groarke, 1996), and therefore visual arguments can also exist. Still others contend that visual representations can function as evidence supporting a claim (Dove, 2012).

Moreover, visualizing an argument does not necessarily mean producing a single image. Arguments may naturally evoke a sequence of mental images — for example, the claim *"Jack missed the shot on goal because he was blinded by sunlight"* is most readily understood as a dynamic situation unfolding across multiple moments. While Champagne and Pietarinen (2020) suggest tackling this with a sequence of visual steps that together convey premises and conclusions, in this work we focus on the single-image approach.

Abstract Statements versus Scene Descriptions Modern text-to-image models are trained on large datasets of image-caption pairs. Captions describe the visible content of an image, i.e., concrete scenes. Consequently, the models primarily learn a mapping between scene descriptions and visual representations (Radford et al., 2021; Segalis et al., 2023). A scene depicts a specific situation, assembled from an actor, an action, specific objects, and visible emotions. An argument, on the other hand, is not a scene description but often a rule, a proverb, or an abstract statement; it is therefore generic (timeless), does not describe a specific situ-

Argument	Identified aspects	Aspect-aware prompt
Sugar makes food more enjoyable.	sugar, enjoyment, eating	A young child is happily eating a bowl of ice cream, savoring every delicious bite. The child's eyes light up with joy as they taste the sweet, creamy goodness of the ice cream, enjoying the sweet, sugary taste that brings them pleasure and happiness.
Dogs require regular outdoor activity which promotes a healthier lifestyle for their owners.	outdoor activity, owner, healthy lifestyle	A playful dog and its owner are enjoying a brisk walk through a park on a sunny day, both of them smiling and full of energy. The dog is wearing a colorful bandana and the owner has a water bottle in their hand. The park is filled with people enjoying the beautiful weather, and there are benches and picnic tables throughout the park for people to relax.
Risk of crime is higher in public transport.	crime, public transport, risk	A crime scene where a violent crime has taken place in a public transport, with police officers investigating the area and taking statements from witnesses.

Table 2: From left to right: The original argument, the identified aspects, and the aspect-aware prompt describing a scene. Aspect identification and prompt generation are performed using LLaMA-3.2-3B-Instruct and Mistral-7B-v0.1, respectively. Using the aspect-aware prompts, the images in Table 1 (second column) were generated.

ation, and lacks identifiable actors.

This difference has direct implications for the concept of "visualizability", which can be defined as follows:

> Visualizability is the degree to which a linguistic description can be directly translated into a concrete visual scene. It depends on referentiality, concreteness, event structure, actor-object relationships, and spatial information.

In our approach we increase the visualizability of the original argument by reducing its ambiguity and subjectivity. For this purpose we identify a small number of aspects (4–8) for a given argument and, based on the two most salient aspects, formulate a scene description as a prompt that conveys the essence of the argument. We refer to our approach as "aspect-aware image generation". Table 2 gives examples for arguments, the identified aspects, and the generated prompts with which the images in Table 1 were generated.

The fact that more concrete descriptions lead to clearer and more coherent visual results is also confirmed by Liu and Chilton (2022).

4 Dataset Construction

To evaluate argumentative image generation, we construct a dataset of generated images for argumentative claims. The claims are sourced from an existing dataset by Heinrich et al. (2025). Each claim is a standalone assertion without supporting premises (we therefore use claim and argument interchangeably throughout), making it well suited for representation in a single image. The claims are argumentative in that they function as enthymemes: arguments whose supporting premises are implicit and reconstructable by the audience from common knowledge (Walton, 2008).

For instance, the claim *"Sugar makes food more enjoyable"* rests on unstated premises such as *sugar makes food sweeter* and *sweeter food is more enjoyable*. This enthymematic structure is characteristic of nearly all claims in our dataset, which range from causal assertions (*"Automation increases work efficiency"*) to evaluative judgments (*"Street art beautifies urban areas"*). It is also what makes image generation challenging: rather than depicting a literal scene description, the generated image must visually convey the implicit reasoning behind the claim—precisely the gap that our aspect-aware approach addresses.

We generate images for each claim using three approaches: Stable Diffusion XL Base 1.0 (Rombach et al., 2022), Stable Diffusion 3.5 Medium (Esser et al., 2024), both of which take the claim directly as input, and our aspect-aware generation approach, which proceeds in three stages. First, five candidate aspects are automatically identified from each argument using LLaMA 3.2 (3B-Instruct) (Grattafiori et al., 2024). Next, a human annotator selects the three aspects that best correspond to the claim. These selected aspects are then passed to Mistral (7B-v0.1) (Jiang et al., 2023) to generate a detailed image prompt, which Stable Diffusion XL Base 1.0 subsequently uses to produce the final image. The choice of models for each stage was guided by preliminary experiments during development. The prompts used for identification and generation are provided in Appendix A. After generation, 127 claims covering 27 topics remain in our dataset.

Interval	Human assessment Aspects	Images
[0.0; 0.5)	429 (11%)	68 (4%)
[0.5; 1.0)	314 (8%)	141 (7%)
[1.0; 1.5)	631 (17%)	582 (31%)
[1.5; 2.0]	2,436 (64%)	1,114 (58%)
Σ	**3,810 (100%)**	**1,905 (100%)**

Table 3: Distribution of the human assessment scores across 1,905 images (127 claims × 3 systems × 5 images). Each image was rated by two annotators on two aspects using a three-point scale (0 = not depicted, 1 = partial, 2 = clear). The two ratings per aspect and the four ratings per image were averaged and assigned to the corresponding interval (first column).

For each argument–generation-approach pair, five images are generated, yielding 1,905 images in total[3].

For each argument, we define two ground-truth aspects that should be identifiable in a generated image to convey the underlying argument. Each image is then checked for the presence of both aspects: every aspect is independently annotated by two annotators using a three-point scale (0–2), where 0 denotes no coverage, 1 denotes partial coverage, and 2 denotes full coverage, resulting in 7,620 annotation points. The final aspect score for an image is computed as the mean of the two annotators' scores, and the final image score is obtained by averaging across the two aspects. Dataset statistics are summarized in Table 3.

Inter-annotator agreement, measured via Cohen's κ at the aspect level, yields $\kappa = 0.34$ — a moderate score that reflects the task's inherent subjectivity. Judging how strongly an image conveys a given argumentative aspect requires interpreting the visual scene in light of the broader claim, rather than applying purely perceptual criteria. Representative annotation examples illustrating this challenge are provided in Appendix B.

5 Experimental Analysis

We conduct two experiments. First, we evaluate our aspect-aware image generation approach against baselines that use the raw argument as a prompt, measuring image quality through human assessment and automated quality indicators. Second, we assess whether the aspects required for aspect-aware generation can be automatically extracted from a given argument.

5.1 Aspect-Aware Image Generation

We evaluate our aspect-aware image generation method against the two Stable Diffusion baselines on our dataset. Representative outputs from all three approaches are shown in Table 1. For each approach, quality indicators are computed per generated image, averaged across the five images generated per argument, and then aggregated across all arguments to obtain the overall score. Results are reported in Table 4. Beyond human assessments, we evaluate image quality using the indicators described below. Since prompt quality directly influences the generated image, two indicators — concreteness and imageability — target the prompt itself; the remaining indicators operate on the generated image. Several of the latter require textual image descriptions, for which we sample five descriptions per image using LLaVA (7B) (Liu et al., 2023) (temperature 0.3) to account for variability in model outputs.

Concreteness Concreteness reflects the degree to which a word refers to a perceptible, physical entity (e.g., *dog* scores high, *justice* scores low), with ratings from 1 (very abstract) to 5 (very concrete). We measure prompt concreteness using the psycholinguistic lexicon of Brysbaert et al. (2014). Since concreteness is defined at the word level, extending it to multi-word prompts is non-trivial (Wu and Smith, 2023). Rather than averaging over all words, we instead aggregate over the five highest-scoring tokens per prompt, focusing the measure on the most content-bearing terms. This is particularly important for aspect-aware prompts, which tend to be longer and thus contain more low-scoring function words (e.g., prepositions) that would otherwise dilute the score.

Imageability Imageability measures how readily a word evokes a mental image (Coltheart, 1981; Wilson, 1988), operationalized by looking up words in a dedicated psycholinguistic lexicon.[4] Although closely related to concreteness, the two constructs can diverge: *infinity*, for instance, may readily evoke a mental image (e.g., of endless space) yet receives a low concreteness score, as it lacks a tangible physical referent. Ratings range from

[3] The images and data were generated during our participation in, and organization of, the Touché shared task (Kiesel et al., 2025; Anand and Heinrich, 2025).

[4] Dataset available at `https://huggingface.co/datasets/StephanAkkerman/MRC-psycholinguistic-database`

System	Human assessment	Prompt quality		Image quality (SBERT)			Image quality (CLIP)	
		Concreteness	Imageability	Interpret-ability	Prompt fidelity	Aspect coverage	Prompt fidelity	Aspect coverage
Ours	0.800 ±0.15	0.924 ±0.04	0.796 ±0.04	0.924 ±0.01	0.774 ±0.05	0.670 ±0.04	0.658 ±0.01	0.619 ±0.01
SD 1.0	0.668 ±0.20	0.594 ±0.12	0.558 ±0.08	0.922 ±0.01	0.668 ±0.05	0.675 ±0.04	0.644 ±0.01	0.620 ±0.01
SD 3.5	0.671 ±0.22	0.594 ±0.12	0.558 ±0.08	0.918 ±0.01	0.676 ±0.05	0.679 ±0.03	0.645 ±0.01	0.618 ±0.01

Table 4: Comparison of our image generation approach using aspect-aware prompts (row "Ours") with Stable Diffusion 1.0 XL and 3.5 Medium using literal argument prompts (rows "SD 1.0" and "SD 3.5"). The second column (Human assessment) shows the achieved ground-truth scores (cf. the overall distribution in Table 3, third column). Columns 3–9 show algorithmically measured quality indicators for both the prompts and the images. For better readability, all scores have been scaled to a range between 0 and 1. The computation of the quality indicators is explained in the text.

1 (low imageability) to 7 (high imageability). As with concreteness, this measure is defined at the word level; to obtain a single imageability score for a prompt, we aggregate over the five highest-scoring tokens.

Interpretability Interpretability measures how consistently an image is described, estimated via a caption-consistency score inspired by Wu and Smith (2023). We compute pairwise cosine similarity over the five generated image descriptions using Sentence-BERT (SBERT, all-MiniLM-L6-v2) (Reimers and Gurevych, 2019), then average the resulting scores. Values range from -1 to 1, where higher scores indicate that independently generated descriptions converge on similar content, suggesting the depicted image is clear and unambiguous.

SBERT Prompt Fidelity Prompt fidelity measures how well the generated image reflects its prompt. Using SBERT, we compute the cosine similarity between the prompt and each of the five image descriptions and average the resulting scores. Values range from -1 to 1, with higher scores indicating closer alignment.

SBERT Aspect Coverage To evaluate whether the image descriptions capture the underlying ground-truth aspects, we compute the cosine similarity between each aspect and every sentence in the description using SBERT, retaining the maximum similarity across sentences. This ensures that an aspect is credited if any single sentence reflects it, rather than diluting the score by averaging over all words. The per-aspect scores are then averaged across both ground-truth aspects and all five image descriptions to yield a single score per image, ranging from -1 to 1.

CLIP Prompt Fidelity and Aspect Coverage Analogous to their SBERT counterparts, these measures use CLIP (ViT-B/32) embeddings (Radford et al., 2021) and cosine similarity, but operate directly on images rather than on textual descriptions, eliminating the need for an intermediate captioning step. Prompt fidelity is computed as the cosine similarity between the generated image embedding and the prompt embedding, following the image–text alignment paradigm of Hessel et al. (2021). Since CLIP truncates text inputs exceeding 77 tokens, longer aspect-aware prompts are partially cut off, which may reduce the measured prompt fidelity for these prompts. Aspect coverage is computed analogously, as the average cosine similarity between the image embedding and the embeddings of each ground-truth aspect.

5.2 Identification of Argument Aspects

We evaluate whether LLMs can automatically recover the ground-truth aspects for a given argumentative claim. To this end, each model is prompted to generate five candidate aspects per claim, and we run each model five times per claim to account for output variability. We experiment with LLaMA 3.2 (3B-Instruct) and GPT-5-Nano (2025-04-07) (Singh et al., 2025). LLaMA is run at temperature 0.3; GPT-5-Nano at its default of 1.0, as the parameter is not user-adjustable. Since a model may return semantically redundant aspects within a single output, we deduplicate the five generated aspects before matching. We report recall over ground-truth aspects — the fraction of the two

Model	Aspect selection	τ=.40	τ=.50	τ=.60	τ=.70
LLaMA 3.2	All	0.909	0.852	0.728	0.595
	Top-2	0.797	0.715	0.607	0.490
GPT-5-Nano	All	0.976	0.964	0.881	0.757
	Top-2	0.859	0.821	0.743	0.617

Table 5: Recall over ground-truth aspects under Hungarian 1-to-1 assignment with SBERT cosine similarity $\geq \tau$, where recall is defined as the fraction of the two ground-truth aspects matched by at least one generated aspect. Results are shown for LLaMA 3.2 (3B-Instruct) and GPT-5-Nano across two aspect selection strategies: "All" considers all five generated aspects, "Top-2" retains only the two most similar to the claim by SBERT cosine similarity. Averaged over five runs.

ground-truth aspects matched by at least one generated aspect — using Hungarian assignment (Kuhn, 1955) to enforce strict one-to-one correspondence between generated and ground-truth aspects. A match is established when SBERT cosine similarity exceeds a threshold. We compare two aspect selection strategies. The first applies Hungarian matching directly against all deduplicated aspects, discarding unmatched ones. The second first reduces the candidate set to the two aspects most similar to the claim by SBERT cosine similarity before matching, constituting a fully automatic setting that requires no human input. Recall across varying thresholds is reported in Table 5.

6 Results and Discussion

Human assessments in Table 4 confirm that the aspect-aware generation approach consistently outperforms both baselines, with Figure 1 providing a holistic view across all quality dimensions. The comparatively high standard deviation of the human scores underscores the subjectivity inherent in evaluating argumentative images, consistent with the moderate inter-annotator agreement reported in Section 4—even though predefined aspects serve as an explicit reference frame for the annotators. Concreteness and imageability scores are notably higher for the aspect-aware approach, suggesting that its prompts more effectively evoke concrete, visually grounded scenes. Regarding interpretability, differences between systems are small, as all three approaches generally produce coherent and interpretable images. This aligns with annotators' qualitative observations: while baseline images are typically understandable, they tend to highlight only a single argumentative aspect while neglecting others, reducing their overall relevance to the underlying claim.

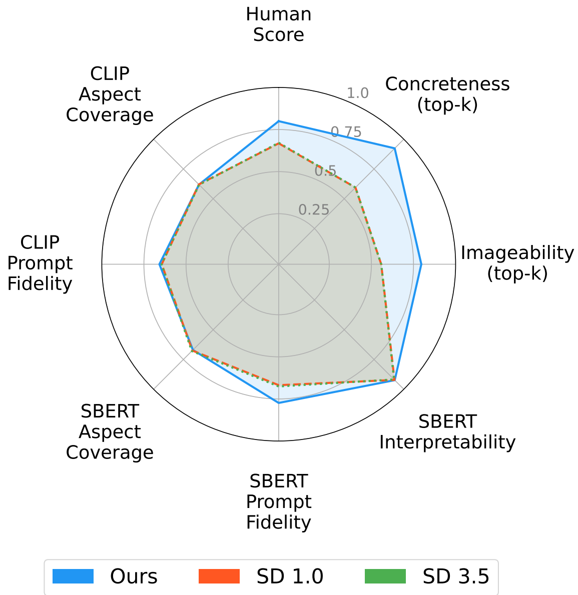


Figure 1: Radar chart comparing aspect-aware generation ("Ours") against Stable Diffusion XL 1.0 ("SD 1.0") and Stable Diffusion 3.5 Medium ("SD 3.5") across eight quality indicators. All scores have been scaled to a range between 0 and 1.

A notable difference is observed for prompt fidelity under both SBERT and CLIP: aspect-aware prompts align considerably more closely with the generated descriptions than raw argumentative claims, confirming that they are more descriptive and yield images that better reflect the intended content. Additionally, CLIP similarity scores tend to cluster within a more compressed range than their SBERT counterparts, a characteristic attributable to the anisotropic nature of CLIP representations (Tyshchuk et al., 2023).

The baseline systems score marginally higher on aspect coverage under SBERT, and under CLIP only Stable Diffusion XL 1.0 achieves a marginally higher score. However, higher coverage does not indicate superior image quality. The reason lies in how the two approaches depict aspects. The baseline uses the claim verbatim as a prompt and tends to depict one concrete aspect literally while neglecting more abstract ones. Consider the claim *"Sugar makes food more enjoyable"* with ground-truth aspects *sugar* and *enjoyable food*: the baseline generates images of candy and sweets whose captions contain words like *candy, sugar, sweet, treats, lollipops* that overlap directly with the *sugar* aspect, yielding high coverage for that aspect, but largely ignore *enjoyable food*. Since the final coverage score averages over both aspects, the high score on the literally depicted aspect can still outweigh the low score on the neglected one. The aspect-aware approach instead constructs a scene description that integrates both aspects indirectly — for instance, a child happily eating ice cream conveys both *sugar* and *enjoyable food*. The resulting captions (*child, ice cream, bowl, joy, dessert*) are distributionally further from the original aspect terms, so the individual coverage scores tend to be lower, and their average falls below that of the baseline despite arguably conveying the argument more effectively. This reveals a limitation of the coverage metric: it rewards literal depiction of aspect terms over holistic scene descriptions that integrate multiple aspects into a coherent visual narrative.

Table 5 shows that the two ground-truth aspects of each claim are consistently recovered among the five LLM-generated aspects. This holds even under Top-2 selection, where retaining only the two aspects most similar to the claim still yields high recall. Recall increases as the similarity threshold τ becomes more permissive, and across all conditions, GPT-5-Nano outperforms LLaMA 3.2. Notably, even at the strictest threshold ($\tau = .70$), GPT-5-Nano retains strong recall scores of 0.757 (All) and 0.617 (Top-2), while LLaMA 3.2 degrades more sharply—suggesting that GPT-5-Nano produces aspects that align more closely with the ground truth.

7 Conclusion

Generating images for arguments poses a fundamental challenge: arguments are typically abstract and rarely describe concrete scenes, while images are inherently ambiguous and open to interpretation. We address this through aspect-aware image generation, which decomposes the process into three explicit steps: aspect identification, prompt construction, and image generation. This grounds abstract argumentative content in a concrete visual scene, improving visualizability. Experiments on a dataset with human relevance assessments show that aspect-aware image generation consistently outperforms baselines that use raw argument text as a prompt. We further demonstrate that aspect-aware prompts are substantially more concrete and imageable than raw argument text, and that LLM-generated aspects can reliably recover the ground-truth aspects defined in the dataset. In principle, an argument could be illustrated without explicitly identifying aspects — for instance, by directly prompting a language model to generate a scene description from the argument. However, the explicit identification of aspects serves two important functions beyond prompt construction: it provides a controllable interface for steering the generation process toward specific argumentative dimensions, and it establishes interpretable criteria against which the resulting images can be evaluated. Future work investigating alternative decomposition strategies or additional aspect types may yield further insights into how argumentative content can be most effectively translated into visual form, while improved evaluation methods could better capture whether generated images convey argumentative intent holistically rather than through literal depiction of individual aspects.

Beyond improving the illustration of arguments, the explicit modeling of argumentative aspects opens further possibilities. By selecting aspects that emphasize specific persuasive goals, images can be tailored to the characteristics and preferences of a target audience. Another promising direction is a more thorough investigation of automated image evaluation. Building on prior work on argument quality (Wachsmuth et al., 2017), future approaches could assess whether generated images effectively convey dimensions such as credibility or emotional appeal, enabling more systematic analysis of how well images support argumentative goals and guiding the optimization of generation approaches.

A further research direction concerns the illustration of more complex arguments — ones that consist of premises and conclusions rather than a single claim, and that therefore require multiple coordinated images. As discussed in Section 3, Champagne and Pietarinen (2020) suggest that understanding certain arguments requires a sequence of visual steps that together convey premises and conclusions. Generating such image sequences introduces the challenge of maintaining semantic and stylistic consistency across images. Techniques for personalization and concept preservation (Ruiz et al., 2023; Gal et al., 2023) may help ensure cross-image coherence, enabling richer forms of visual argumentation. In such settings, the generation process could be decomposed into a sequence of subgoals, with intermediate visual outputs combined post hoc to construct a coherent argumentative narrative (Chern et al., 2025).

8 Limitations and Ethical Considerations

The subjective nature of images makes standardized evaluation inherently difficult, and our annotation does not assess dimensions such as emotional impact, perceived realism, or persuasive power — all of which are highly relevant to images used in argumentative contexts. Furthermore, the evaluation considers aspects in isolation rather than examining whether multiple aspects are coherently integrated within a single image. Additionally, while our pipeline uses specific models for each stage based on preliminary experiments, a systematic evaluation of how different language and image generation models affect each stage of the pipeline — from aspect identification to prompt construction to image generation — remains an open question.

Regarding image generation, potential biases in generated images must also be acknowledged: generative models are trained on large-scale datasets that reflect cultural and societal influences, and are therefore far from neutral (Morales et al., 2025). More broadly, images function as powerful communicative tools: they are often perceived as direct reflections of reality, lending them an appearance of inherent truth and credibility (Grancea, 2017). Synthetic or manipulated images can therefore destabilize established notions of visual evidence and trust (Momeni, 2025), particularly when deployed to frame events or reinforce specific interpretations of complex issues. As generative models make the production of persuasive visual content increasingly accessible, the generation and use of argument-supporting images demand careful and responsible application.

Acknowledgments

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

References

Sarah K. Alhabeeb and Amal A. Al-Shargabi. 2024. Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction. *IEEE Access*, 12:24412–24427.

Sharat Anand and Maximilian Heinrich. 2025. Hanuman at Touché: Image Generation with Argument-Aspect Fusion. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025*, volume 4038 of *CEUR Workshop Proceedings*, pages 4571–4579. CEUR-WS.org.

Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Ameneh Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, and Shuaiwen Leon Song. 2025. Renaissance: A survey into AI text-to-image generation in the era of large model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3):2212–2231.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barrière, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. Overview of Touché 2023: Argument and Causal Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, volume 14163 of *Lecture Notes in Computer Science*, pages 507–530. Springer.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of touché 2022: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 311–336. Springer.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Marc Champagne and Ahti-Veikko Pietarinen. 2020. Why images cannot be arguments, but moving ones might. *Argumentation*, 34(2):207–236.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. 2025. Thinking with generated images. arXiv 2505.22525.

Jiwan Chung, Sungjae Lee, Minseo Kim, Seungju Han, Ashkan Yousefpour, Jack Hessel, and Youngjae Yu. 2024. Selective vision is the challenge for visual reasoning: A benchmark for visual argument understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2423–2451. Association for Computational Linguistics.

James M. Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational Psychology Review*, 3(3):149–210.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 33A(4):497–505.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pages 70–98. Association for Computational Linguistics.

Ian J. Dove. 2012. On Images as Evidence and Arguments. In Frans H. van Eemeren and Bart Garssen, editors, *Topical Themes in Argumentation Theory: Twenty Exploratory Studies*, Argumentation Library, pages 223–238. Springer Netherlands, Dordrecht.

Finis Dunaway. 2018. Images, Emotions, Politics. *Modern American History*, 1(3):369–376.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR / OpenReview.net.

David Freedberg. 1989. *The Power of Images: Studies in the History and Theory of Response*. University of Chicago Press, Chicago.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. 2025. Infogen: Generating complex statistical infographics from documents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 20552–20570. Association for Computational Linguistics.

Ioana Grancea. 2017. Types of Visual Arguments. *Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric*, 15(2):16–34.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Leo Groarke. 1996. Logic, art and argument. *Informal Logic*, 18:105.

Leo Groarke. 2024. Informal Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. Metaphysics Research Lab, Stanford University.

Maximilian Heinrich, Johannes Kiesel, Moritz Wolter, Martin Potthast, and Benno Stein. 2025. Touché25: Image Retrieval and Generation for Arguments.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics.

Tamás Janusko, Aaron Kämpf, Denis Keiling, Jessica Knick, David Schäfer, and Maik Thiele. 2024. HTW-DIL at Touché: Multimodal Dense Information Retrieval for Arguments. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 3401–3406. CEUR-WS.org.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ralph H. Johnson. 2003. Why "visual arguments" aren't arguments. In *OSSA Conference Archive*.

Dana Khartabil, Christopher Collins, S. Wells, Benjamin Bach, and Jessie Kennedy. 2021. Design and evaluation of visualization techniques to facilitate argument exploration. *Comput. Graph. Forum*, 40(6):447–465.

Dora Kiesel. 2025. *Effective Visual Analytics for Exploring Argumentation and Deliberation in Text Corpora*. Ph.D. thesis, Bauhaus-Universität Weimar, Weimar.

Johannes Kiesel, Çagri Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Sharat Anand, Tomaz Erjavec, Matthias Hagen, Matyás Kopp, Nikola Ljubesic, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevicius, Harrisen Scells, Moritz Wolter, Ines Zelch, and 2 others. 2025. Overview of Touché 2025: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain,* *September 9-12, 2025, Proceedings*, volume 16089 of *Lecture Notes in Computer Science*, pages 486–508. Springer.

Johannes Kiesel, Çagri Çöltekin, Maximilian Heinrich, Maik Fröbe, Milad Alshomary, Bertrand De Longueville, Tomaz Erjavec, Nicolas Handke, Matyás Kopp, Nikola Ljubesic, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevicius, Theresa Reitis-Münstermann, Mario Scharfbillig, Nicolas Stefanovitch, Henning Wachsmuth, Martin Potthast, and Benno Stein. 2024. Overview of Touché 2024: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II*, volume 14959 of *Lecture Notes in Computer Science*, pages 308–332. Springer.

Jens E. Kjeldsen. 2015. The Rhetoric of Thick Representation: How Pictures Render the Importance and Strength of an Argument Salient. *Argumentation*, 29(2):197–215.

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Vivian Liu and Lydia B. Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 384:1–384:23. ACM.

Max Moebius, Maximilian Enderling, and Sarah T. Bachinger. 2023. Jean-Luc Picard at Touché 2023: Comparing Image Generation, Stance Detection and Feature Matching for Image Retrieval for Arguments. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 3111–3118. CEUR-WS.org.

Mina Momeni. 2025. Artificial intelligence and political deepfakes: Shaping citizen perceptions through misinformation. *Journal of Creative Communications*, 20(1):41–56.

Sergio Morales, Robert Clarisó, and Jordi Cabot. 2025. Imagebite: A framework for evaluating representational harms in text-to-image models. In *4th IEEE/ACM International Conference on AI Engineering - Software Engineering for AI, CAIN 2025, Ottawa, ON, Canada, April 27-28, 2025*, pages 95–106. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22500–22510. IEEE.

Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. 2023. A picture is worth a thousand words: principled recaptioning improves image generation. arXiv 2310.16656.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ 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.

Kirill Tyshchuk, Polina Karpikova, Andrew Spiridonov, Anastasiia Prutianova, Anton Razzhigaev, and Alexander Panchenko. 2023. On isotropy of multimodal embeddings. *Inf.*, 14(7):392.

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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 176–187. Association for Computational Linguistics.

Douglas Neil Walton. 2008. The three bases for the enthymeme: A dialogical theory. *Journal of Applied Logic*, 6(3):361–379.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Si Wu and David Smith. 2023. Composition and deformance: Measuring imageability with a text-to-image model. In *Proceedings of the 5th Workshop on Narrative Understanding*, pages 106–117, Toronto, Canada. Association for Computational Linguistics.

A Prompts

```
List exactly 5 unique aspects (as single words or short phrases) from the following text.
Format them as a numbered list:

“{text}”

Only output the aspect names. Do not include explanations or sentences.
```

Figure 2: Aspect Generation Prompt to identify aspects out of argumentative claim (used with LLaMA 3.2 3B-Instruct).

```
You are a creative assistant tasked with generating descriptive prompts for image generation.
Given an argument and its aspects, craft a detailed, vivid prompt that combines them into a
single, cohesive description suitable for generating an image. The prompt should be rich in
detail, incorporating the argument and all aspects naturally.

Argument: {argument}
Aspects: {aspects}
```

Figure 3: Image-Prompt Generation Prompt to combine Arguments and Aspects into a descriptive prompt (used with Mistral 7B-v0.1).

B Annotation Examples

Argument	Aspect-aware prompt	Argument as literal prompt	
Violent video games promote aggression.			
Aspect: gaming Aspect: aggression	(2,2) (2,2)	(0,2) (2,2)	(2,2) (2,2)
Cultural traditions often rely on the use of animals.			
Aspect: animal Aspect: tradition	(2,2) (2,2)	(2,2) (2,2)	(1,0) (2,2)
	Stable Diffusion 1.0 XL	Stable Diffusion 1.0 XL	Stable Diffusion 3.5 Medium

Table 6: Examples of image generation for argumentative prompts. The left column lists an argument and the aspects that should be visually depicted. Images are generated either using an aspect-aware prompt or by using the argument itself as the prompt. Below each image, tuples denote the ratings assigned by Annotator 1 and Annotator 2 for the corresponding aspect, on a three-point scale (0 = not depicted, 1 = partially depicted, 2 = clearly depicted). Using the argument as a literal prompt can lead to ambiguous interpretations. For example, the Stable Diffusion 1.0 image for *Violent video games promote aggression* received scores of (0,2) for the *gaming* aspect, since the scene can be interpreted as real-world violence rather than a video game context.

Do We Need Large Models for Argument Classification? Revisiting the Role of Model Compression

Filip Gampel[1], Rafał Olszowski[1], Marcin Pietroń[2],

[1]Faculty of Humanities, AGH University of Krakow,
[2]Faculty of Computer Science, Electronics, and Telecommunications, AGH University of Krakow
Correspondence: rolszowski@agh.edu.pl

Abstract

Large language models have improved argument mining substantially, but the associated computational cost complicates deployment, replication, and systematic comparison. We examine how much compression an open-source large language model can tolerate before argument classification quality degrades. Using gpt-oss-20b as the base model, we study pruning with Wanda and post-training quantization under a zero-shot prompting setup. We evaluate compressed variants on three argument-mining resources, namely UKP, Args.me, and ARIES, and contrast their behavior with general language-model benchmarks. The results show a consistent pattern: moderate pruning preserves most of the original performance on argument classification, whereas activation quantization causes larger and more systematic drops. The findings suggest that argument classification is more compression-tolerant than general-purpose evaluation suites, but only up to a point, and they should not be interpreted as evidence that aggressive compression is universally safe. We therefore position compression as a practical way to reduce model cost for argument analysis, while emphasizing that claims about efficiency gains must distinguish between preserved predictive quality and realized runtime speedups.

1 Introduction

Large language models (LLMs) now deliver strong results on many argument-mining tasks, including stance classification, claim detection, and relation prediction. This progress has come with rapidly increasing model size, memory requirements, and inference cost, which makes evaluation harder to reproduce and limits deployment in resource-constrained settings. For argument mining, this tension is especially relevant because many practical uses require running inference over large collections rather than a handful of carefully selected examples.

Despite the practical importance of efficiency, compression has received far less attention in argument mining than in general-purpose LLM evaluation. Most published work focuses either on better prompting strategies or on stronger base models, while relatively little is known about whether argument classification requires the full capacity of a modern open-source LLM. This gap matters because tasks that look fragile on broad reasoning benchmarks may still be stable for narrower classification problems.

Recent studies of LLM-based argument classification also suggest that performance gains do not come only from scaling parameter counts. Pietroń et al. show that compact or moderately sized models can remain competitive when paired with carefully designed inference procedures and prompting strategies (Pietroń et al., 2024, 2025). That observation motivates our question from a different angle: if argument classification is already less dependent on sheer scale than many other LLM tasks, compression may be especially promising here.

This paper studies that question directly. We start from gpt-oss-20b and apply pruning and post-training quantization, then measure the effect of these interventions on both standard LLM benchmarks and three argument-mining datasets. Our goal is not to claim that compression always improves argument classification, but rather to characterize where performance remains stable and where it begins to degrade.

Our contributions are threefold. First, we provide a focused empirical analysis of compression for argument classification across datasets with different label spaces and discourse structures. Second, we show that moderate sparsification is substantially less harmful on argument classification than on several general benchmarks. Third, we identify the limits of this robustness: higher sparsity and especially activation quantization lead to noticeable deterioration, and gains in accuracy do

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 86–93
July 3, 2026 ©2026 Association for Computational Linguistics

not always translate into gains in macro-F1.

2 Related Work

Argument mining has evolved from feature-based pipelines to neural and pretrained language-model approaches (Mochales and Moens, 2011; Lippi and Torroni, 2016; Lawrence and Reed, 2020). Representative datasets cover cross-topic argument identification (Stab et al., 2018), large-scale argument search (Ajjour et al., 2019), and persuasive discussion analysis (Chakrabarty et al., 2019). More recent systems build on pretrained encoders and LLM prompting, which has made zero-shot and few-shot argument classification increasingly competitive (Devlin et al., 2019; Wei et al., 2022; Bar-Haim et al., 2017).

Within argument classification specifically, recent work has explored how smaller open models can be strengthened through prompting, auxiliary reasoning steps, and post-processing rather than by relying exclusively on ever larger base models (Pietroń et al., 2024, 2025). Our paper is complementary to that line of research: instead of improving a fixed model with additional reasoning machinery, we ask how much of the model can be removed or quantized before performance deteriorates.

Model compression has a long history in neural NLP and deep learning more broadly. Pruning removes parameters judged unimportant, while quantization reduces numerical precision to lower memory and computation requirements (Han et al., 2015; Hinton et al., 2015). For LLMs, methods such as Wanda and GPTQ show that large models often contain substantial redundancy, at least on broad benchmark suites (Sun et al., 2024; Frantar et al., 2023).

What remains underexplored is whether the same compression behavior holds for argument mining. Argument classification involves structured semantic distinctions, but it is still a constrained prediction problem with small output spaces. That combination makes it plausible that such tasks require less effective model capacity than open-ended generation or broad reasoning benchmarks.

3 Method

3.1 Compression Setup

Let the base model be parameterized by weights

$$\Theta = \{W_1, \ldots, W_L\}, \tag{1}$$

where each $W_l \in \mathbb{R}^{d_l \times k_l}$ denotes a learned linear operator in layer l. Compression aims to reduce the effective storage and computation associated with Θ while retaining predictive quality.

For pruning, we apply a binary mask $M_l \in \{0,1\}^{d_l \times k_l}$ to each layer,

$$\bar{W}_l = M_l \odot W_l, \tag{2}$$

where $\odot$ denotes element-wise multiplication. In our setup, pruning is unstructured rather than structured: individual weights are set to zero according to Wanda's activation-aware criterion (Sun et al., 2024). This distinction matters because unstructured sparsity may preserve accuracy without automatically yielding wall-clock speedups unless the inference stack supports sparse kernels.

For post-training quantization, weights are mapped to a lower-precision representation. Using a uniform b-bit quantizer, a weight matrix W is approximated as

$$Q_b(W) = \Delta \cdot \mathrm{round}\left(\frac{W}{\Delta}\right), \tag{3}$$

where Δ is a scale determined by the dynamic range of W. The quantized weights are then

$$\bar{W} = Q_b(W). \tag{4}$$

In the experiments below, we consider 8-bit weight quantization and a more aggressive 8-bit weight-plus-activation setting. The round-to-nearest (RTN) approach is used. Each layer is quantized independently, and quantization parameters are determined individually for each channel within a single layer. In case of activations, quantization is performed dynamically (parameters are determined separately for each token).

3.2 Base Model and Pipeline

Figure 1 summarizes the pipeline. We use gpt-oss-20b as the base model for all experiments because it is a popular open-source model and has already shown strong argument-classification performance in prior work (Pietroń et al., 2025). Starting from the original checkpoint, we produce pruned variants at multiple sparsity levels using Wanda. We then evaluate these models either without additional quantization, with 8-bit weight quantization, or with 8-bit quantization for both weights and activations. No task-specific fine-tuning is performed.

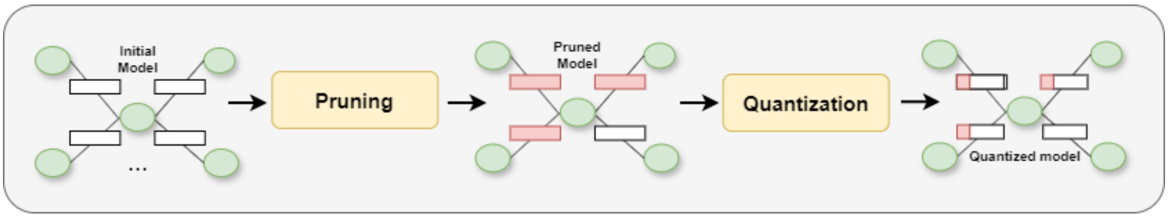


Figure 1: Overview of the experimental pipeline. Starting from gpt-oss-20b, we apply pruning and post-training quantization, evaluate the resulting variants on general and argument-mining benchmarks, and compare how compression affects each task family.

4 Experimental Setup

4.1 Datasets

We evaluate compression on three argument-mining resources chosen to cover distinct formulations of the task.

The UKP corpus (Stab et al., 2018) is a cross-topic argument-mining benchmark built from online comments on controversial issues. We treat it as a three-way classification problem with labels *For*, *Against*, and *No Argument*.

Args.me (Ajjour et al., 2019) contains arguments paired with debated theses from online portals such as Debatewise, Debatepedia, and iDebate. We use it as a binary stance classification task with *For* and *Against* labels.

ARIES (Gemechu et al., 2024) evaluates argumentative relation identification between pairs of argumentative discourse units. We cast it as a three-class classification problem with *Support*, *Attack*, and *No Relation*.

For UKP and Args.me, very long instances were filtered and the evaluation used stratified subsamples of 2,000 instances per subset due computational constraints. Sampling details are summarized in Appendix B.

4.2 Evaluation Protocol

All experiments use fixed task prompts and zero-shot inference. The prompt templates are reported in Appendix B, and the output-parsing summary is reported in Appendix D. For argument-mining datasets, we report accuracy and macro-F1. For the general LLM benchmarks, we report accuracy.

Compression is evaluated at pruning levels of 10%, 20%, 30%, 40%, and 50%. We compare three settings: pruning alone, pruning with 8-bit weight quantization, and pruning with 8-bit weight-plus-activation quantization. The general-benchmark results are included to contextualize how compression affects broad language-model capability relative to argument classification.

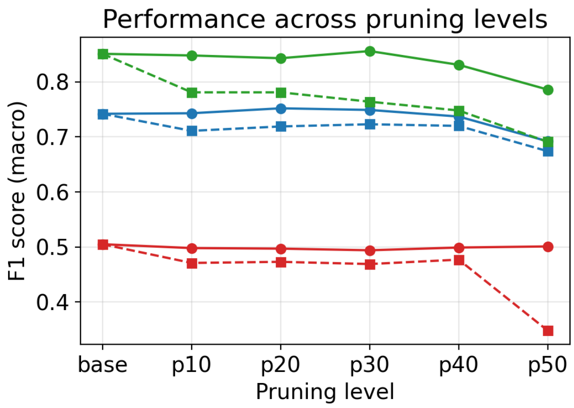


Figure 2: Performance across pruning levels and quantization settings on the argument-mining datasets. Green denotes Args.me, blue UKP, and red ARIES. Solid lines correspond to 8-bit weight quantization; dashed lines additionally quantize activations.

5 Results and Discussion

5.1 General Benchmarks

Table 1 shows that the base model remains broadly stable under moderate compression. Up to 30% pruning with 8-bit weight quantization, most benchmark scores fluctuate only slightly around the baseline. Beyond that point, deterioration becomes more visible, especially on ARC-Easy and ARC-Challenge. The 8-bit weight-plus-activation configuration is clearly less robust even at 40% sparsity.

This pattern is a useful reference for the argument-mining results: once pruning becomes too aggressive, the model's general capability profile degrades.

5.2 Argument Classification

Figure 2 and the detailed results in Appendix A show that argument classification is more compression-tolerant than the general benchmarks, although this robustness is not uniform across

Table 1: Accuracy of pruned gpt-oss-20b variants on general language-model benchmarks.

Model	ARC-Easy	ARC-Challenge	HellaSwag	PIQA	Lambada	Winogrande
base	0.82	0.45	0.58	0.79	0.08	0.67
p10-8w	0.81	0.44	0.57	0.78	0.08	0.69
p20-8w	0.81	0.44	0.60	0.80	0.09	0.67
p30-8w	0.82	0.45	0.58	0.79	0.08	0.65
p40-8w	0.80	0.40	0.58	0.80	0.09	0.68
p50-8w	0.77	0.39	0.60	0.80	0.05	0.69
p40-8w8a	0.73	0.44	0.57	0.73	0.09	0.68

datasets or metrics.

For pruning without additional quantization, UKP is remarkably stable up to 40% sparsity: accuracy stays within 0.005 of the baseline and macro-F1 remains effectively unchanged. Args.me is slightly more sensitive, but the degradation remains small through 30% sparsity and becomes substantial only at 40–50%. ARIES behaves differently. Accuracy rises slightly at higher sparsity, peaking at 0.626 for 50% pruning, but macro-F1 stays essentially flat around 0.50. This means the apparent gain on ARIES should be interpreted cautiously: it likely reflects class-distribution effects rather than a clear improvement in balanced predictive quality.

Adding 8-bit weight quantization preserves the same overall picture. UKP and Args.me remain competitive through moderate sparsity, and ARIES again shows little change in macro-F1 despite small accuracy movements. Weight-only quantization therefore does not materially worsen the compression profile already induced by pruning.

The more aggressive 8-bit weight-plus-activation setting is different. Here, performance drops are systematic across all three datasets. UKP and Args.me lose around four to ten accuracy points depending on sparsity, and ARIES suffers the strongest macro-F1 deterioration at 50% sparsity. The practical conclusion is straightforward: moderate pruning is often acceptable for argument classification, but activation quantization is riskier in this setting.

6 Limitations

The study has several limitations that should temper the conclusions. It covers a single base model family, so the observed robustness may not transfer directly to other architectures. We also evaluate predictive quality rather than direct latency, throughput, or memory savings, which matters because unstructured sparsity does not automatically translate into faster inference on standard hardware. Finally, UKP and Args.me are evaluated on subsamples and the experiments rely on fixed prompts and deterministic answer parsing, which increases uncertainty around small differences.

7 Conclusion

We revisited the role of model compression in argument classification by evaluating pruned and quantized variants of gpt-oss-20b on UKP, Args.me, and ARIES. The central result is that argument classification remains stable under moderate pruning and weight-only quantization even when general LLM benchmarks begin to deteriorate. At the same time, the evidence does not support stronger claims that aggressive compression is universally safe; activation quantization, in particular, introduces substantial risk.

Taken together, the results support a pragmatic conclusion: for argument classification, moderate compression is often a defensible efficiency strategy, but it should be validated with task-specific metrics and runtime measurements.

8 Future works

Several extensions follow naturally from the present study. One direction is task-specific fine-tuning, which would help determine whether compression remains effective once the model is adapted more closely to argument-classification datasets. A second direction is to evaluate multi-prompt inference strategies in which predictions from several prompt formulations are aggregated, as this may improve robustness when individual prompts are sensitive to wording. A third direction is to design prompts that target recurring error sources more directly, particularly contrastive discourse structures, multi-faceted arguments, multiple negations, and cases in which referential alignment is lost.

9 Funding

This research was funded by Narodowe Centrum Nauki (National Science Centre, Republic of Poland), the research grant UMO-2023/49/B/HS5/01379, "Argument Mining: Public Debate Models and Algorithmically-Assisted Argument Extraction" ("Argument mining: Modele debat publicznych i wspomagana algorytmicznie ekstrakcja argumentów"). We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018546.

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data acquisition for argument search: The args.me corpus. In C. Benzmüller and H. Stuckenschmidt, editors, *KI 2019: Advances in Artificial Intelligence*, volume 11793 of *Lecture Notes in Computer Science*. Springer, Cham.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. 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 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *International Conference on Learning Representations*.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. ARIES: A general benchmark for argument relation identification. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

John Lawrence and Chris Reed. 2020. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marcin Pietroń, Filip Gampel, Jakub Gomułka, Andrzej Tomski, and Rafał Olszowski. 2025. A Comprehensive Study of LLM-Based Argument Classification: from Llama through DeepSeek to GPT-5.2. In preprint.

Marcin Pietroń, Rafał Olszowski, and Jakub Gomułka. 2024. Efficient argument classification with compact language models and chatgpt-4 refinements. In N.T. et al. Nguyen, editor, *Computational Collective Intelligence. ICCCI 2024*, volume 14810 of *Lecture Notes in Computer Science*. Springer, Cham.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 3664–3674. Association for Computational Linguistics.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964.

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 *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Table 2: Performance of pruned models without additional quantization.

	UKP		Args.me		ARIES	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
base	0.781	0.742	0.857	0.851	0.605	0.505
p10	0.780	0.744	0.848	0.842	0.601	0.499
p20	0.785	0.751	0.852	0.845	0.597	0.498
p30	0.783	0.747	0.855	0.848	0.591	0.494
p40	0.782	0.744	0.840	0.832	0.610	0.505
p50	0.750	0.694	0.807	0.791	0.626	0.503

Table 3: Performance of pruned models with 8-bit weight quantization.

	UKP		Args.me		ARIES	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
base	0.781	0.742	0.857	0.851	0.605	0.505
p10	0.779	0.743	0.854	0.848	0.601	0.498
p20	0.785	0.752	0.850	0.843	0.596	0.497
p30	0.784	0.749	0.863	0.856	0.589	0.494
p40	0.776	0.737	0.839	0.831	0.608	0.499
p50	0.749	0.692	0.800	0.786	0.624	0.501

A Detailed results

Detailed calculation results are presented in tables 2, 3, 4.

B Prompts

For UKP, `{topic}` is either one of "abortion", "cloning", "death penalty","legalisation of marijuana","stricter gun laws","minimum wage","nuclear energy", "school uniforms", depending on the dataset. In all other cases, data is taken directly and literally from the dataset.

UKP: Is the sentence: "`{sentence}`" an argument for or against `{topic}`, or is it no argument? Return one of the expressions: “For”, “Against” or “No argument”, without any additional commentary.

Argsme: "Is the sentence: "`{sentence}`" an argument for or against `{thesis}`? Return one of the expressions: “For” or “Against”, without any additional commentary."

ARIES: Given the two propositions:
Proposition 1: "`{prop1}`"
Proposition 2: "`{prop2}`"
What is the argumentative relation between Proposition 1 and Proposition 2? Return one of the expressions: "Support", "Attack" or "No Relation", without any additional commentary.

Table 4: Performance of pruned models with 8-bit weight and 8-bit activation quantization.

	UKP		Args.me		ARIES	
Model	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
base	0.781	0.742	0.857	0.851	0.605	0.505
p10	0.742	0.711	0.787	0.781	0.585	0.471
p20	0.751	0.719	0.787	0.781	0.588	0.473
p30	0.757	0.723	0.771	0.764	0.581	0.469
p40	0.740	0.710	0.760	0.748	0.589	0.477
p50	0.710	0.654	0.721	0.691	0.611	0.348

C Datasets

For UKP and Args.me, we performed an initial screening of all datasets to crop very long records (>2000 characters in the `argument/sentence` field). Due to computational restraints, for each dataset we used subsamples of 2000 records. These were generated randomly from the full sets, ensuring that the original class imbalance is preserved. Class counts can be found below:

Table 5: UKP class counts

Dataset	For	Against	NoArg	Total
abortion	346	418	1236	2000
cloning	465	552	983	2000
death	250	609	1141	2000
gun	471	398	1131	2000
marijuana	474	506	1020	2000
nuclear	339	476	1185	2000
school	362	485	1153	2000
wage	466	446	1088	2000

Table 6: Args.me class counts

Dataset	For	Against	Total
debatepedia	1490	510	2000
debatewise	1179	821	2000
idebate	988	1012	2000

D Answer Parsing

To extract the final classification from the model outputs, we applied regular expressions (Regex) designed to capture the expected label formats while ignoring extraneous text (e.g., "The answer is..."). Below are the regex patterns used for parsing the cleaned model outputs (reasoning, whitespace and punctuation removed, converted to lowercase):

UKP: `r'^\b(for|against|noargument)\b|\b(for|against|noargument)\b$'`
Argsme: `r'^\b(for|against)\b|\b(for|against)\b$'`
ARIES: `r'^\b(norelation|support|attack)\b$'`

E Experimental Setup and Hyperparameters

All local inferences were conducted on the Athena supercomputer at the Academic Computer Centre Cyfronet AGH. The computations were performed on nodes equipped with 8 $\times$ NVIDIA A100 GPUs (40GB VRAM each).

The models were deployed using the vLLM library. We utilized the default vLLM sampling parameters, with the exception of temperature, which was set to 0.6 for all models. The maximum number of new tokens generated was set to 4096. Other hyperparameters such as reasoning effort were also left to the default values.

A Neural Approach to Fine-Grained Argumentation Strategy Classification with Emotion and Moral Value Lexicons across Multiple Domains

Mohammad Yeghaneh Abkenar[1,3], **Weixing Wang**[2,3], **Manfred Stede**[3] **and Julia Romberg**[4]
[1]Innovation Department, Bundesdruckerei Gruppe GmbH Berlin [2]Hasso Plattner Institute
[3]University of Potsdam [4]GESIS - Leibniz Institute for the Social Sciences
yeghanehabkenar@uni-potsdam.de, weixing.wang@hpi.de,
stede@uni-potsdam.de, julia.romberg@gesis.org

Abstract

Fine-grained argumentation mining goes beyond coarse-grained distinctions such as claim and premise, by delving deeper into the underlying strategies employed (e.g., the use of facts or values to persuade the audience). Despite the advancements brought about by pre-trained language models, the task remains challenging. We investigate whether auxiliary knowledge such as emotion and moral value lexicon features can improve the classification of fine-grained argumentation strategies. Our Neural Flair Transformer Classifier (NFTC), in its base form, fine-tunes a transformer-based document encoder (RoBERTa) for end-to-end argument component classification. Evaluated across four corpora from diverse domains spanning public participation, persuasive forums, product reviews, and student essays, NFTC consistently outperforms majority-voting and Qwen2.5-7B baselines, achieving competitive performance on all datasets. Moreover, gains are observed against a fine-tuned LLaMA-3-8B-Instruct model, regarded in prior work as a leading approach. Injecting additional knowledge into NFTC yields mixed effects: emotion and moral value features provide consistent gains in product reviews and persuasive forums, but not in the other two domains. Our findings suggest that the utility of subjective knowledge is domain and schema dependent, and that knowledge enrichment beyond standard pre-training can meaningfully complement transformer-based models for fine-grained argumentation mining. We provide all resources—including code, the preprocessed corpora, and model architecture—to enable other researchers to build upon our work.[1]

1 Introduction

One of the key tasks in argument mining is the *computational assessment of the function of argumentative units* in natural text. Traditionally, this task has centered on coarse-grained distinctions such as claim (a statement that expresses a specific point or conclusion) and premise (a statement that provides support or justification for a claim) (Palau and Moens, 2009; Liebeck et al., 2016; Stab and Gurevych, 2017; Daxenberger et al., 2017).

While claim/premise classification has been fundamental for argument mining, many applications can benefit from a *more fine-grained analysis that reveals the underlying strategies employed to make claims and premises persuasive for the audience* (Park and Cardie, 2014; Hidey et al., 2017; Dusmanu et al., 2017; Park and Cardie, 2018; Schaefer et al., 2023). Examples include the analysis of persuasion strategies in news editorials (Al-Khatib et al., 2016), or recommending arguments that follow a specific argumentation strategy, which can be used, for example, in debates (Rinott et al., 2015).

The more recent approaches to computationally classifying fine-grained argumentation strategies have put a strong emphasis on the use of pre-trained language models that are subsequently fine-tuned on task-specific data (Schaefer and Stede, 2022; Schaefer et al., 2023; Cabessa et al., 2025). This shift has certainly improved performance, but seems not yet sufficient to accommodate the complex nature of argumentation.

One reason might be the need for additional knowledge beyond the information encoded in the model during the pre-training and fine-tuning phases. Such knowledge can be "any kind of normative information that is considered to be relevant for solving a task at hand and that is not given as task input itself" (Lauscher et al., 2022). In other computational argumentation tasks, we can already observe that such knowledge enrichment can provide further performance improvements, such as in uncovering implicit information (Becker et al., 2020), in audience-specific claim generation (Alshomary et al., 2021), or in stance detection (Abkenar et al., 2026).

[1]The code for the experiments can be found here: **FineAM**.

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 94–106
July 3, 2026 ©2026 Association for Computational Linguistics

Our main research question in this paper is therefore **whether, and to what extent, additional knowledge can improve the classification of fine-grained argumentation strategies**. We focus on two sources of information. Our first hypothesis is that *emotionality* can provide useful additional information, as emotional appeal is considered a potential factor affecting the rhetorical effectiveness of arguments (Wachsmuth et al., 2017; Vecchi et al., 2021). Our second hypothesis is that the notion of *moral values* can also provide useful additional information, since references to shared values or moral principles can be effective in persuading audiences in argumentative discourse (Alshomary and Wachsmuth, 2021; Kiesel et al., 2022).

Recent work has shown that argumentation mining models tend to rely on shortcuts and corpora-specific features rather than learning generalizable argumentation properties or subtask, stressing the need to evaluate across different domains (Feger et al., 2025). Motivated by this finding, we test our knowledge-enriched neural models across multiple domains to examine whether lexicon-based features (emotion and moral values) contribute to more robust and generalizable fine-grained argumentation strategy classification.

Our contributions are as follows:

- We evaluate the impact of emotionality signals for fine-grained argumentation strategy classification, using a knowledge-enriched transformer architecture that integrates emotion lexicons (Abkenar et al., 2026).
- We extend the model architecture towards the incorporation of moral value signals.
- We evaluate the approach across four heterogeneous datasets spanning public participation, persuasive forums, product reviews, and essays.
- We find that incorporating emotion and moral value features improves model performance in half of the cases, while leading to a decrease in fine-grained argumentation strategy classification performance for the other half.

2 Related Work

While coarse-grained classification (e.g., into claim and premise) has been a major focus area in argument mining, the more fine-grained look at the argumentation strategies that authors pursue has received comparatively less attention (Schaefer et al., 2023). This is unfortunate, as fine-grained labels often provide richer and more detailed information about argumentative structures.

2.1 Fine-grained Argumentation Schemas

What exactly constitutes a fine-grained argumentation strategy is interpreted diversely and to some extent depends on the goal of a study as well as the domain of the data. Al-Khatib et al. (2016, 2017) target persuasion strategies in news editorials using a schema of *common ground*, *assumption*, *testimony*, *statistics*, *anecdote*, and *other*. In the domain of persuasive essay writing, Carlile et al. (2018) categorize argumentation strategies into the claim types *fact*, *value* and *policy*, which is in line with distinctions made in argumentation theory (cf. (Eggs, 2000)). As premise types, they use *common knowledge*, *real example*, *invented instance*, *warrant*, *statistics*, *testimony*, *definition* and *analogy*. As shown in Figure 1, Schaefer et al. (2023) modify this schema with respect to the premise types, resulting in *testimony*, *statistics*, *hypothetical-instance*, *real-example* and *common-ground*.

To inform a debating agent with Wikipedia knowledge, Rinott et al. (2015) extracts premise types as *study*, *expert*, and *anecdote*. More interactive data sources have also been explored. Dusmanu et al. (2017) and Schaefer and Stede (2022) annotate tweets, either as *factual* or *opinionated*, or by classifying a claim as *unverifiable* or *verifiable* and an evidence as *reason*, *external* or *internal*. Park and Cardie (2018) evaluate fine-grained argumentation strategies (*policy*, *value*, *fact*, *testimony*, *reference*) with the goal of measuring the evaluability of arguments in the context of online public participation platforms. The same schema has also been applied to assess the helpfulness of product reviews (Chen et al., 2022). Park and Cardie (2014) operate on online public participation discussions as well, providing us with a different approach that emphasizes the factuality of propositions. Claims are distinguished as *unverifiable*, *verifiable non-experiential*, and *verifiable experiential*, while premise strategy types are *reason*, *evidence*, and *optional evidence*. Other discussion forum sources include idebate.org, distinguishing supporting arguments as *study*, *factual*, *opinion*, and *reasoning* (Hua and Wang, 2017), and Change My View: (Hidey et al., 2017) use a rather unique

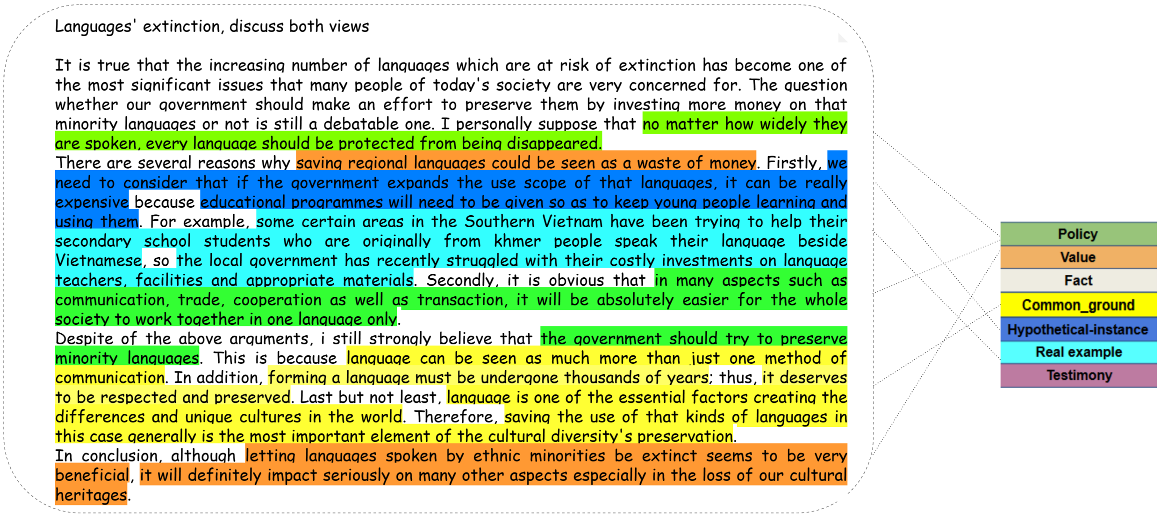


Figure 1: Example of a fine-grained argumentation strategy schema, taken from Schaefer et al. (2023) (AAE-Ext corpus): an argumentative essay annotated with five out of a total of seven fine-grained semantic types (policy, value, fact, common ground, hypothetical instance, real example, testimony). Some labels do not appear in this essay but occur in others.

set of claim types, namely *interpretation*, *evaluation*, *agreement*, and *disagreement*. For premises, Hidey et al. make use of Aristotle's *logos*, *pathos*, and *ethos*, as do a few other works (Habernal and Gurevych, 2017; Carlile et al., 2018).

2.2 Neural Approaches for Fine-grained Argumentation Strategy Classification

When developing approaches to computationally predicting fine-grained argumentation strategies, recent works have demonstrated the dominance of neural methods. Typically, these approaches rely on existing pre-trained language models that are subsequently fine-tuned on task-specific data, such as BERT and RoBERTa (Schaefer and Stede, 2022; Schaefer et al., 2023). In addition, large language models such as LLaMA-3, Gemma-2, Mistral, Phi-3, and Qwen-2 have been utilized for this purpose (Cabessa et al., 2025). A somewhat distinct approach is employed by Chen et al. (2022), who extract argumentative features as input for neural models. Their feature extraction follows Morio et al. (2020)'s method of feeding word features (surface, part-of-speech tags, GloVe and ELMo vectors) into a BiLSTM to predict strategy types.

While these methods achieve promising results, it has been argued that computational models for argumentation require the integration of additional knowledge to adequately capture the complexity and subjective nature of argumentative discourse (Lauscher et al., 2022). We try to fill this gap to a certain extent by revisiting some previously collected datasets, performing preprocessing to prepare corpora suitable for neural methods, and leveraging the injection of **additional knowledge about emotionality and moral values**.

More broadly, fine-grained argumentation mining provides a promising setting for exploring neural models and capturing subtle argumentative phenomena that coarse-grained frameworks may overlook (Ren et al., 2025).

2.3 Emotionality and Moral Values as Signals for Argumentation Strategies

Subjectivity is an integral part of argumentation, with expressions of sentiment, emotion, and affect serving as important signals (Lauscher et al., 2022). Early work provides evidence that such features can be helpful in fine-grained argumentation strategy classification (Rinott et al., 2015; Levy et al., 2014; Dusmanu et al., 2017; Hua and Wang, 2017).

Emotional appeal can affect the rhetorical effectiveness of arguments (Wachsmuth et al., 2017; Vecchi et al., 2021), which is central to the fine-grained study of argumentation properties that make claims and premises persuasive. In particular, research on argument quality often identifies emotion as a key dimension of relevance (Benlamine et al., 2015, 2017; Ziegenbein et al., 2023; Greschner and Klinger, 2025; Quensel et al., 2025; Chen et al., 2026). Pre-neural approaches on fine-grained argumentation strategy classification thus regularly incorporated sentiment and emotion features (Levy et al., 2014; Dusmanu et al., 2017; Hua and Wang, 2017), in contrast to neural approaches.

References to shared values or moral principles are equally subjective cues (van der Meer et al., 2023; Homayounirad et al., 2025) that can be used as persuasive strategies in arguments (Bench-Capon, 2003). Among the first that connected human values with argument mining was Kiesel et al. (2022), focusing on predicting the values behind arguments. Further studies followed up on this (Jafari et al., 2024; Zhang et al., 2024; Senthilkumar et al., 2025) or explored the role of values in argument generation and rewriting (Alshomary and Wachsmuth, 2021; Shahid et al., 2026). To the best of our knowledge, moral values have not yet been exploited to inform our task at hand.

In sum, these insights motivate our goal to **complement promising neural methods with the subjective signals of emotionality and moral values**.

3 Methodology

We adopt the classification framework recently introduced by Abkenar et al. (2026) for stance classification, combining transformer-based text representations with additional knowledge-driven features. The framework allows external signals to be incorporated alongside contextual embeddings, which makes it suitable for our research question.

3.1 Model Architecture

We first detail the original model architecture, which incorporates emotion features. We then outline how we adapt the architecture in order to include moral value features.

Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ denote an argumentative text sequence input consisting of n tokens. The goal of the model is to predict a fine-grained argumentation label $y \in \mathcal{Y}$, where $\mathcal{Y}$ represents the set of dataset-specific fine-grained argumentation categories.

Transformer Encoder. We encode the input text using a pre-trained transformer-based document encoder. The input sequence is formatted as

`[CLS] Argument:` $\mathbf{x}$ `[SEP]`

The transformer produces contextualized token representations

$$\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\} \in \mathbb{R}^{n \times d},$$

where d is the hidden dimensionality of the encoder. The representation corresponding to the `[CLS]` token, $\mathbf{h}_{\text{CLS}} \in \mathbb{R}^d$, is used as the document-level representation of the argument.

The transformer encoder is fine-tuned end-to-end during training. All models are implemented using the Flair NLP framework[2] with HuggingFace transformer encoders, which we refer to as the Neural Flair Transformer Classifier (NFTC).

Lexicon- and Knowledge-based Features. To enrich the semantic representation with psychologically grounded signals, we incorporate two types of auxiliary features derived from external resources.

- **Emotion features (eNRC).** Emotion information is extracted using the best-performing variant of the extended NRC Emotion Lexicon (eNRC)[3], as introduced by Abkenar et al. (2026). The eNRC expands prior versions of the NRC emotion lexicon (Zad et al., 2021; Mohammad and Turney, 2013) by incorporating additional emotive and affective vocabulary not previously captured. Each lexicon entry is associated with zero or more emotion labels based on Plutchik's Wheel of Emotions (Plutchik, 2001), one of the most influential models in emotion research. This model captures eight primary emotions: *joy*, *trust*, *fear*, *surprise*, *sadness*, *disgust*, *anger*, and *anticipation*.

  For each input text, this produces an emotion feature vector

  $$\mathbf{f}_{\text{emo}} \in \mathbb{R}^8.$$

  Emotion features are constructed by aggregating token-level eNRC scores across the input text, where each emotion category frequency is normalized by the total number of emotionally matched tokens, yielding an 8-dimensional document-level feature vector.

- **Moral foundation features (MoralBERT).** We additionally model moral framing using predictions from MoralBERT[4] classifiers trained on the Moral Foundations Theory (Haidt et al., 2017; Graham et al., 2011). This theoretical framework assumes that human moral reasoning is based on a set of innate, universal moral domains (so-called "foundations"). It organizes morality into ten core

[2] `https://github.com/flairNLP/flair`
[3] https://github.com/Pioannid/eNRC/tree/main
[4] https://github.com/vjosapreniqi/MoralBERT

Corpora	Domain	Texts	Units	Fine-grained types
CMV	persuasive forum	121 (threads)	4612	policy, value, fact, testimony, rhetorical statement
CDCP	public participation	731 (comments)	4931	policy, value, fact, testimony, reference
AAE-Ext	students' persuasive essays	402 (essays)	6089	policy, value, fact, testimony, statistics, real-example, hypothetical-instance, common-ground, other
AM^2	product reviews	878 (reviews)	6126	policy, value, fact, testimony, reference

Table 1: Corpus statistics of the four preprocessed corpora, annotated with fine-grained argumentation strategies, that are used in our evaluation.

values: *care*, *harm*, *fairness*, *cheating*, *loyalty*, *betrayal*, *authority*, *subversion*, *purity*, and *degradation*.

For each input, MoralBERT outputs a 10-dimensional feature vector,

$$\mathbf{f}_{\text{moral}} \in \mathbb{R}^{10},$$

where each dimension corresponds to one moral foundation category. The value of each entry represents the model-estimated presence or strength of the corresponding moral foundation in the text, normalized to the interval $(0, 1)$. Higher values indicate stronger evidence that the text expresses the corresponding moral foundation.

Feature Integration. The auxiliary features are integrated with the contextual representation via feature-wise concatenation. Given the transformer document embedding $\mathbf{h}_{\text{CLS}}$, the final representation is defined as

$$\mathbf{h}_{\text{final}} = \begin{cases} \mathbf{h}_{\text{CLS}} & \text{(no auxiliary features)} \\ [\mathbf{h}_{\text{CLS}}; \mathbf{f}_{\text{emo}}] & \text{(emotion features)} \\ [\mathbf{h}_{\text{CLS}}; \mathbf{f}_{\text{moral}}] & \text{(moral features)} \end{cases} \quad (1)$$

where $[\cdot;\cdot]$ denotes vector concatenation. Following the approach of Bravo-Marquez et al. (2019), and Abkenar et al. (2026), we have done a simple concatenation of features to be able to compare the effect of both features on the results.

Classification. The resulting document representation $\mathbf{h}_{\text{final}}$ is passed to a linear classification layer that predicts the label distribution over $\mathcal{Y}$. The model is trained using standard cross-entropy loss.

3.2 Corpora and Statistics

In our evaluation, we focus on four English datasets from different domains. These corpora were selected from established argumentation datasets and subjected to extensive pre-processing steps:

- **ChangeMyView (CMV)** (Morio et al., 2019).[5] ChangeMyView is a Reddit subforum dedicated to changing users' views through persuasive argumentation. The corresponding dataset contains 4612 discussion turns, each annotated as *policy*, *value*, *fact*, *testimony*, or *rhetorical statement*. The CMV corpus is provided in Brat[6] annotation format. Each thread contains three posts; an original post, a positive reply, and a negative reply and both inner-post and inter-post relations from which we extract argumentation units and their semantical components types.

- **Cornell eRulemaking Corpus (CDCP)** (Park and Cardie, 2018).[7] CDPC was among the first datasets to adopt a more fine-grained annotation schema, aiming to model argumentative structures as they occur in real-world scenarios. Collected from a US public participation effort, the dataset classifies 4931 argumentative units into *policy*, *value*, *fact*, *testimony*, and *reference*. The CDCP corpus is provided in JSONL format. We extract proposition texts and their semantic components types.

[5] https://katfuji.lab.tuat.ac.jp/nlp_datasets/
[6] https://brat.nlplab.org/
[7] https://huggingface.co/datasets/DFKI-SLT/cdcp

- **Argument Annotated Essays Corpus (AAE-Ext)** (Schaefer et al., 2023).[8] AAE-Ext builds upon the AAE corpus of persuasive essays (Stab and Gurevych, 2017) by introducing an additional annotation layer. It features the largest set of labels in our collection, including *policy*, *value*, *fact*, *testimony*, *statistics*, *real-example*, *hypothetical-instance*, *common-ground*, and *other*. The AAE-Ext corpus is provided in Brat annotation format and consists of 402 essays. We extract the 6089 argument units and their labels *MajorClaim*, *Claim*, and *Premise*. We then integrate the extended version to extract the fine-grained semantic component types.
- **AMazon Argument Mining Corpus (AM^2)** (Chen et al., 2022).[9] 6126 text units from Amazon reviews were categorized according to a fine-grained schema of *policy*, *value*, *fact*, *testimony*, and *reference*. Each entry contains a review with a list of propositions, each annotated with an identifier, argumentation type, text, reasons, and evidence. We extract the proposition texts and their corresponding type labels as input-output pairs for classification.

Table 1 presents the statistics for each corpus, including domain, size, and fine-grained argumentation strategy types.

3.3 Experimental Setup

NFTC. Motivated by the strong results of *RoBERTa* among other encoder models in argument component classification (Schaefer et al., 2023), we fine-tune *RoBERTa-base* for sequence classification as the transformer encoder used in our NFTC approach. Training is conducted for 10 epochs. We use a fixed learning rate of 5×10^{-5} and a mini-batch size of 64.

We evaluate several model configurations of the proposed NFTC classifier. The base model (referred to as NFTC hereinafter) uses only the transformer encoder. To examine the impact of knowledge-enhanced features, we additionally incorporate emotion features derived by means of the eNRC emotion lexicon. We refer to this model variant as NFTC + eNRC. We also evaluate a variant that integrates moral foundation features predicted by MoralBERT (NFTC + MoralBERT). In all cases, the auxiliary feature vectors are concatenated with the transformer [CLS] representation before classification.

Baselines. We compare our three NFTC variants to various baselines. We run a simple majority-class baseline (MajC) and the zero-shot Qwen2.5-7B (Hui et al., 2024) LLM baseline used by Abkenar et al. (2026) on our four evaluation datasets. For Qwen2.5-7B, we use this prompt: *You are an expert text classification model. Task: Read the input text and assign exactly ONE label from the list. Labels: [...] Instructions: - Output ONLY the label text from the list above. - Do NOT output explanations or anything else. Text: [...] Answer (one label only):* Model temperature was set to 0.0.

We further include previously reported results on these datasets, which are taken verbatim from the respective research papers. For the AM^2 dataset, we refer to Chen et al. (2022), who report the performance of a model based on BiLSTM encoders. For CDCP, we resort to the results from Cao (2023), who introduce AutoAM, a multi-task learning model that integrates BERT with a specialized argumentation attention mechanism. For CMV, we reference the results obtained by the structured SVM model in Galassi et al. (2018). For AAE-Ext, however, a direct comparison with prior baseline results is not possible due to differences in the evaluation methodology.

In addition, we compare our approach to LLama-3-8B-Instruct, which was found to be the excelling LLM in argument component classification (Cabessa et al., 2025). Fine-tuning was performed with LLaMA-Factory in SFT mode using QLoRA. We train for 5 epochs on a batch size of 8 and a learning rate of $5e^{-5}$ is used. We fine-tuned in 4-bit quantized form, following prior evidence that this introduces negligible performance differences (Cabessa et al., 2025).

General Settings. We use a 5-fold cross-validation setting in all experiments. The folds remain the same. We have used a seeding method in our publicly accessible code, so the results are fully reproducible. All neural experiments were conducted on a single NVIDIA RTX 3090 GPU.

All models are evaluated on the four corpora (AM^2, CDCP, CMV, and AAE-Ext). Following prior work, model performance is reported as the mean macro F_1 score across the cross-validation folds, due to class imbalance in several datasets. We also report standard deviation across the five folds.

[8] https://github.com/discourse-lab/arg-essays-semantic-types

[9] https://facultystaff.richmond.edu/˜jpark/

	AM^2	CDCP	CMV	AAE-Ext
MajC	0.151 (± 0.001)	0.122 (± 0.005)	0.132 (± 0.001)	0.050 (± 0.022)
Qwen2.5-7B	0.629 (± 0.006)	0.724 (± 0.005)	0.685 (± 0.005)	0.389 (± 0.010)
Further Baselines	0.496	0.846	0.735	-
LLaMA-3-8B-Instruct	0.762 (± 0.000)	0.845 (± 0.000)	0.796 (± 0.000)	0.522 (± 0.005)
NFTC (ours)	0.839 (± 0.008)	**0.856 (± 0.010)**	0.834 (± 0.008)	**0.589 (± 0.010)**
NFTC + eNRC	0.858 (± 0.007)	0.821 (± 0.009)	0.840 (± 0.007)	0.567 (± 0.010)
NFTC + MoralBert	**0.861 (± 0.005)**	0.821 (± 0.006)	**0.849 (± 0.005)**	0.463 (± 0.190)

Table 2: Macro F_1 comparison of our Neural Flair Transformer Classifier (NFTC), with the combined eNRC, and moral features variants, the majority-class baseline (MajC), Qwen2.5-7B, LLaMA-3-8B-Instruct, and prior baselines on the corpora. Prior baseline results are taken from Chen et al. (AM^2), Cao (CDCP), and Galassi et al. (CMV).

4 Results and Discussion

Table 2 presents the performance of the different models on the four datasets.

Across datasets, our proposed NFTC classifier demonstrates strong performance and consistent with respect to the standard deviation. On CDCP and AAE-Ext, the base NFTC model achieves the best results, with macro F_1 scores of 0.856 and 0.589, respectively. On AM^2 and CMV, incorporating MoralBERT features yields the highest performance, achieving scores of 0.861 and 0.849. These results substantially outperform the other models.

4.1 Baseline and LLM as a Judge

The majority-class baseline performs poorly across all datasets, achieving macro F_1 scores between 0.050 and 0.151. This confirms the difficulty of fine-grained argumentation strategy classification, particularly for datasets with substantial class imbalance or a larger number of categories such as AAE-Ext (details on the corpora distribution are provided in Appendix A). All neural network-based models substantially outperform this baseline.

Motivated by these findings, we employ Qwen2.5-7B due to its high performance, specialized reasoning capabilities, and efficiency for deployment. The Qwen2.5-7B model provides stronger performance, but remains consistently below the results achieved by our fine-tuned transformer models.

Overall, our NFTC model consistently surpasses both the majority baseline and the LLM baseline across all datasets, indicating that task-specific fine-tuning remains highly effective for fine-grained argumentation strategy classification.

4.2 Prior Models for Fine-grained Argument Classification

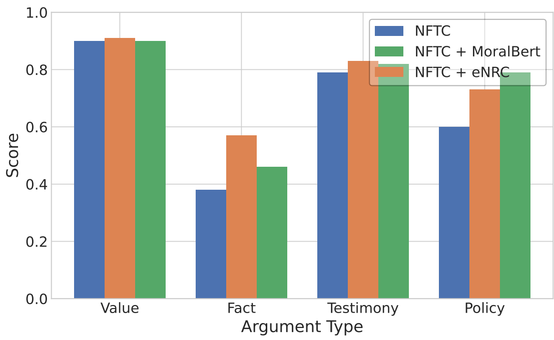
We first compare against results reported on the respective datasets in prior work. As shown in Table 2, for AM^2, NFTC clearly outperforms the previously reported results by Chen et al. (2022). We also obtain better results on CDCP, with 0.856 a slightly higher F_1 score than the previously reported 0.845 of Park and Cardie (2018). For CMV, we achieve improved results with all NFTC variants. Although NFTC yields the best results for AAE-Ext overall, a direct comparison with prior baseline results is not possible due to differences in the evaluation methodology.

Next, we compare our NFTC models to a fine-tuned LLaMA-3-8B-Instruct, which is considered the current state-of-the-art model for argument component classification (Cabessa et al., 2025). While the original authors evaluated their model only on CDCP as a fine-grained argumentation schema, we extend the evaluation to include three additional datasets. Our results demonstrate that even a fine-tuned LLaMA-3-8B-Instruct model does not outperform NFTC variants: for AM^2 and CMV, NFTC consistently achieves better performance, and for CDCP[10] and AAE-Ext, at least one NFTC variant surpasses LLaMA-3-8B-Instruct.

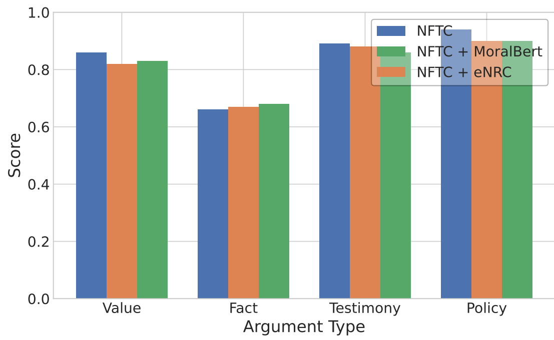
4.3 Effect of Knowledge-Enriched Features

NFTC variants perform best on all datasets. We further evaluate the effect of incorporating external

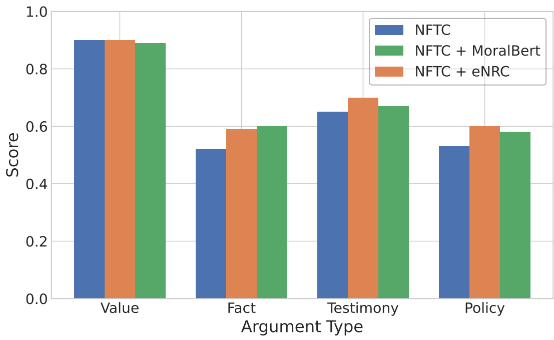
[10] Cabessa et al. (2025) report a higher mean macro F_1 score of 0.873, compared to our experimental result. However, their evaluation was conducted on a predefined single test split, while we employ a 5-fold cross-validation setup. We argue that our approach provides a more robust estimate of model performance, as it mitigates the risk of results being skewed by the selection of a particular test set.



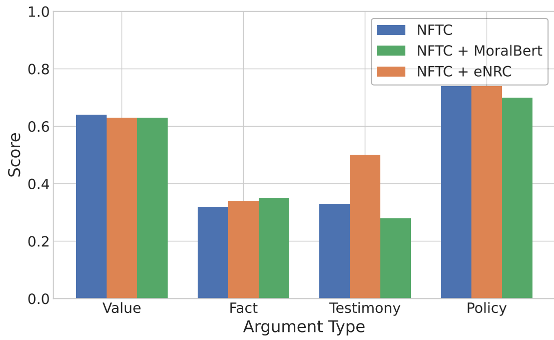
(a) AM^2 dataset



(b) CDCP dataset



(c) CMV dataset



(d) AAE-Ext dataset

Figure 2: Breakdown of the macro F_1 scores for NFTC in its base form and for the injection of additional knowledge, moral values (NFTC + MoralBERT) and emotion (NFT + eNRC). We report the performance for the overlapping categories, i.e., value, fact, testimony and policy, for all four datasets.

knowledge through emotion features (eNRC) and moral foundation features (MoralBERT). The results indicate that the usefulness of these auxiliary features is dataset-dependent. MoralBERT features improve performance on AM^2 and CMV, suggesting that moral framing contributes to identifying argumentation strategies in product reviews and online discussions. Emotion features also provide improvements, but these are smaller on AM^2 and CMV compared to the NFTC base model.

Moreover, both auxiliary knowledge enrichments lead to a performance drop on the other two datasets, CDCP and AAE-Ext. This finding is particularly interesting when comparing CMV and CDCP, which might be expected to be more similar: one being a general discussion forum and the other focused on political topics. However, a notable difference is likely the target audience: while CMV consists of discussions among users, public participation submissions in CDCP are more directly addressed to officials. For AAE-Ext, the results may also reflect the broader and more heterogeneous annotation schema used in this dataset, where lexical signals alone may provide limited additional information.

Knowledge-based features can complement transformer representations, but their effectiveness depends on the domain, the function of the text, and, presumably, the annotation schema.

4.4 Category-Level Analysis

In order to gain further insights into the impact of emotion and moral value enrichment on the indiviual argumentation strategy categories, Figure 2 presents the category-level performance of the four dataset. The results show that the *value*, *testimony* and *policy* categories are classified with relatively high macro F_1 scores across all model variants, while the *fact* category exhibits lower scores overall.

Notably, the integration of emotion and moral features consistently improves the classification of *facts* across all datasets, and to some extent also enhances the prediction of the *testimony* and *policy* categories. This suggests that these strategies may rely more strongly on affective or normative cues captured by the auxiliary lexicon-based features. In the case of the objectively normed category of *facts*, a possible explanation is that the absence of emotional or moral value cues provides informative signals to the model. The only category that remains unaffected by model enrichment is *values*, which is particularly interesting given that this category typically reflects highly subjective propositions. Future research is needed to determine whether this is due to a mismatch between the theoretical foundations of our auxiliary features and the *values* category, or if other factors might be contributing to this effect.

5 Conclusion

In this paper, we investigated whether auxiliary knowledge about emotionality and moral values can improve fine-grained argumentation strategy classification. We evaluated our Neural Flair Transformer Classifier (NFTC) across four diverse corpora spanning public participation, persuasive forums, product reviews, and student essays. Our results demonstrate that NFTC consistently outperforms both the majority-voting baseline and Qwen2.5-7B, and achieves competitive performance on all datasets where prior results are available. Furthermore, gains are noted against the fine-tuned LLaMA-3-8B-Instruct model.

The injection of auxiliary knowledge yields mixed effects: MoralBERT features provide consistent gains on AM2 and CMV, while emotion features via eNRC prove more beneficial for AAE-Ext. This suggests that the utility of subjective knowledge is domain and schema dependent, with moral framing being more informative in interactive discourse settings and emotion signals better suited to richer, multi-class annotation schemes. Taken together, our findings support the hypothesis that knowledge enrichment beyond standard pre-training and fine-tuning can meaningfully complement transformer-based models for fine-grained argumentation mining.

Future work should explore more targeted integration strategies — such as attention-based feature fusion (Dai et al., 2021) or gated feature fusion (Li et al., 2020) — and investigate whether other sources of normative knowledge can further close the gap in harder, multi-class settings such as AAE-Ext, as well as extending this approach to relation classification.

Limitations

While our NFTC approach works well from a performance perspective, the improvement brought by the auxiliary features is small. Future work must explore the benefits of auxiliary knowledge in more depth, such as through more complex feature integration methods as well different emotion lexicon features such as SenticNet (Cambria et al., 2016). Additionally, the prompt used for our Qwen2.5-7B baseline was relatively generic and could have been optimized. The results may be further improved by employing prompt engineering approaches.

Ethical Considerations

We are fully aware that systems based on emotion recognition and sentiment analysis can be facilitators of enormous progress, but also enablers of great harm. We therefore strongly advice user of such systems to follow established instructions and ethical guidelines, such as ethics sheets (Mohammad, 2022) and data-sheets for datasets (Gebru et al., 2021). Moreover, emotional expression varies significantly across cultures, ethnic groups, and demographics, as well as moral values. This must be carefully considered, especially when using such systems to support policy-making (i.e., the evalution of public participation).

Acknowledgments

We thank our colleagues in Innovations department of the Bundesdruckerei GmbH, Hasso Plattner Institute and Leibniz Institute for the Social Sciences (GESIS) for providing us with the opportunity to freely work on our research topics. Thank you for fostering an environment that encourages innovation and academic growth. The authors also sincerely thank the anonymous reviewers for their thoughtful recommendations that significantly improved this paper.

References

Mohammad Yeghaneh Abkenar, Weixing Wang, Manfred Stede, Mark A. Finlayson, Davide Picca, and Panagiotis Ioannidis. 2026. Improving neural argumentative stance classification in controversial topics with emotion-lexicon features. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9678–9691, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark. Association for Computational Linguistics.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan. The COLING 2016 Organizing Committee.

Milad Alshomary, Wei-Fan Chen, Timon Gurcke, and Henning Wachsmuth. 2021. Belief-based generation of argumentative claims. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 224–233, Online. Association for Computational Linguistics.

Milad Alshomary and Henning Wachsmuth. 2021. Toward audience-aware argument generation. *Patterns*, 2(6):100253.

Maria Becker, Ioana Hulpuş, Juri Opitz, Debjit Paul, Jonathan Kobbe, Heiner Stuckenschmidt, and Anette Frank. 2020. Explaining arguments with background knowledge. *Datenbank Spektrum*, 20:131–141.

Trevor J. M. Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Mohamed S. Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive argumentation and emotions: An empirical evaluation with users. In *Human-Computer Interaction. User Interface Design, Development and Multimodality*, pages 659–671, Cham. Springer International Publishing.

Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. Emotions in Argumentation: an Empirical Evaluation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, pages 156–163, Buenos Aires, Argentina.

Felipe Bravo-Marquez, Eibe Frank, Bernhard Pfahringer, and Saif M. Mohammad. 2019. Affectivetweets: a weka package for analyzing affect in tweets. *Journal of Machine Learning Research*, 20(92):1–6.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Björn Schuller. 2016. Senticnet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2666–2677.

Lang Cao. 2023. Autoam: An end-to-end neural model for automatic and universal argument mining. In *International Conference on Advanced Data Mining and Applications*, pages 517–531. Springer.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Yanran Chen, Lynn Greschner, Roman Klinger, Michael Klenk, and Steffen Eger. 2026. Emotionally charged, logically blurred: AI-driven emotional framing impairs human fallacy detection. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6709–6732, Rabat, Morocco. Association for Computational Linguistics.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8922, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. 2021. Attentional feature fusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3560–3569.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Ekkehard Eggs. 2000. Vertextungsmuster Argumentation: Logische Grundlagen. In Klaus Brinker, editor, *Text- und Gesprächslinguistik*, volume 16 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 397–414. Walter de Gruyter, Berlin.

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.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2018. Argumentative link prediction using residual networks and multi-objective learning. In *Proceedings of the 5th Workshop on Argument Mining*, pages 1–10.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366.

Lynn Greschner and Roman Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 628–646, Albuquerque, USA. Association for Computational Linguistics.

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

Jonathan Haidt, P Ditto, J Graham, R Iyer, S Koleva, M Motyl, G Sherman, and S Wojcik. 2017. Moral foundations theory. *Social Theorists of Morality*, page 261.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Homayounirad, Enrico Liscio, Tong Wang, Catholijn M Jonker, and Luciano Cavalcante Siebert. 2025. Will annotators disagree? identifying subjectivity in value-laden arguments. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15237–15252, Suzhou, China. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208, Vancouver, Canada. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Amir Reza Jafari, Praboda Rajapaksha, Reza Farahbakhsh, Guanlin Li, and Noel Crespi. 2024. Unveiling human values: Analyzing emotions behind arguments. *Entropy*, 26(4).

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, Shaohua Tan, and Kuiyuan Yang. 2020. Gated fully fusion for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11418–11425.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics.

Saif Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and predicting online persuasion strategy with elementary units. 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 6274–6279, Hong Kong, China. Association for Computational Linguistics.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2020. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3259–3266, Online. Association for Computational Linguistics.

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*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. Investigating subjective factors of argument strength: Storytelling, emotions, and hedging. In *Proceedings of the 12th Argument mining Workshop*, pages 126–139, Vienna, Austria. Association for Computational Linguistics.

Yupei Ren, Xinyi Zhou, Ning Zhang, Shangqing Zhao, Man Lan, and Xiaopeng Bai. 2025. Towards comprehensive argument analysis in education: Dataset, tasks, and method. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14215–14231.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

Robin Schaefer, René Knaebel, and Manfred Stede. 2023. Towards fine-grained argumentation strategy analysis in persuasive essays. In *Proceedings of the 10th Workshop on Argument Mining*, pages 76–88, Singapore. Association for Computational Linguistics.

Robin Schaefer and Manfred Stede. 2022. GerCCT: An annotated corpus for mining arguments in German tweets on climate change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6121–6130, Marseille, France. European Language Resources Association.

Rithik Appachi Senthilkumar, Amir Homayounirad, and Luciano Cavalcante Siebert. 2025. Leveraging large language models to identify the values behind arguments. In *Value Engineering in Artificial Intelligence*, pages 87–103, Cham. Springer Nature Switzerland.

Farhana Shahid, Stella Zhang, and Aditya Vashistha. 2026. Llms homogenize values in constructive arguments on value-laden topics. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA. Association for Computing Machinery.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Michiel van der Meer, Piek Vossen, Catholijn M. Jonker, and Pradeep K. Murukannaiah. 2023. Do differences in values influence disagreements in online discussions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore. Association for Computational Linguistics.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

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.

Samira Zad, Joshuan Jimenez, and Mark Finlayson. 2021. Hell hath no fury? correcting bias in the NRC emotion lexicon. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 102–113, Online. Association for Computational Linguistics.

He Zhang, Alina Landowska, and Katarzyna Budzynska. 2024. Detection and analysis of moral values in argumentation. In *Value Engineering in Artificial Intelligence*, pages 114–141, Cham. Springer Nature Switzerland.

Timon Ziegenbein, Shahbaz Syed, Felix Lange, Martin Potthast, and Henning Wachsmuth. 2023. Modeling appropriate language in argumentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4344–4363, Toronto, Canada. Association for Computational Linguistics.

A Appendix

Dataset	**Label**	**Count**	**(%)**
AM^2 (6126)	value	3752	61.25
	testimony	1965	32.08
	fact	281	4.59
	policy	124	2.02
	reference	4	0.07
CDCP (4931)	value	2177	44.15
	testimony	1118	22.67
	policy	815	16.53
	fact	789	16.00
	reference	32	0.65
CMV (4727)	value	3134	66.30
	rhetorical_statement	727	15.38
	testimony	354	7.49
	fact	257	5.44
	policy	140	2.96
	major_claim	115	2.43
AAE-Ext (6089)	common_ground	1774	29.13
	value	1502	24.67
	hypothetical_instance	917	15.06
	real_example	717	11.78
	fact	411	6.75
	statistics	400	6.57
	policy	344	5.65
	testimony	22	0.36
	other	2	0.03

Table 3: Label distribution across datasets. Counts and percentages are reported for each class.

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
July 3, 2026 ©2026 Association for Computational Linguistics

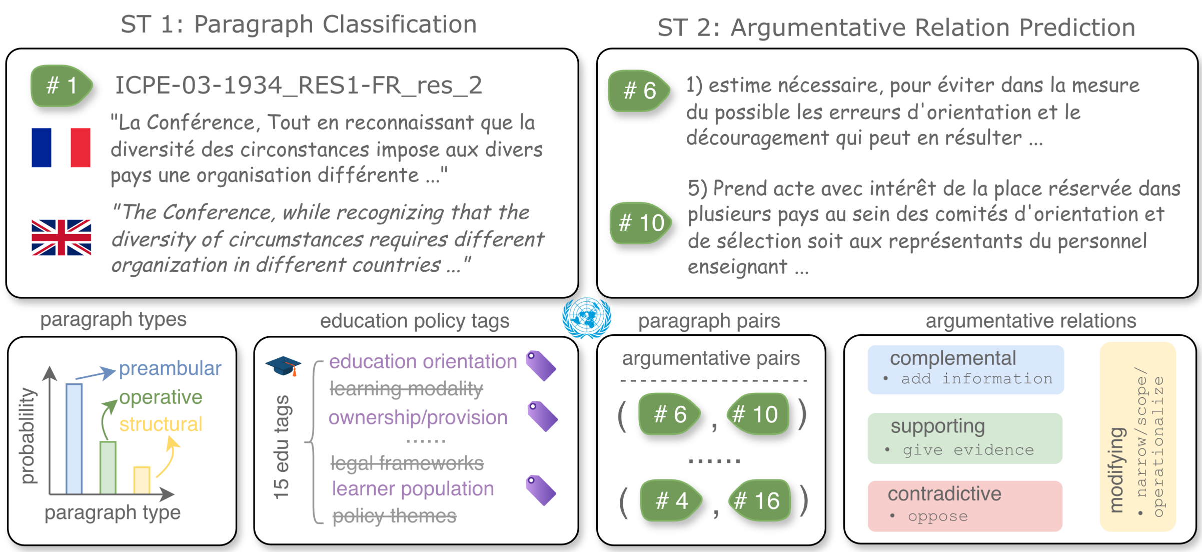


Figure 1: Overview of the UZH Shared Task 2026 on Reconstructing the Reasoning in United Nations Resolutions.

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

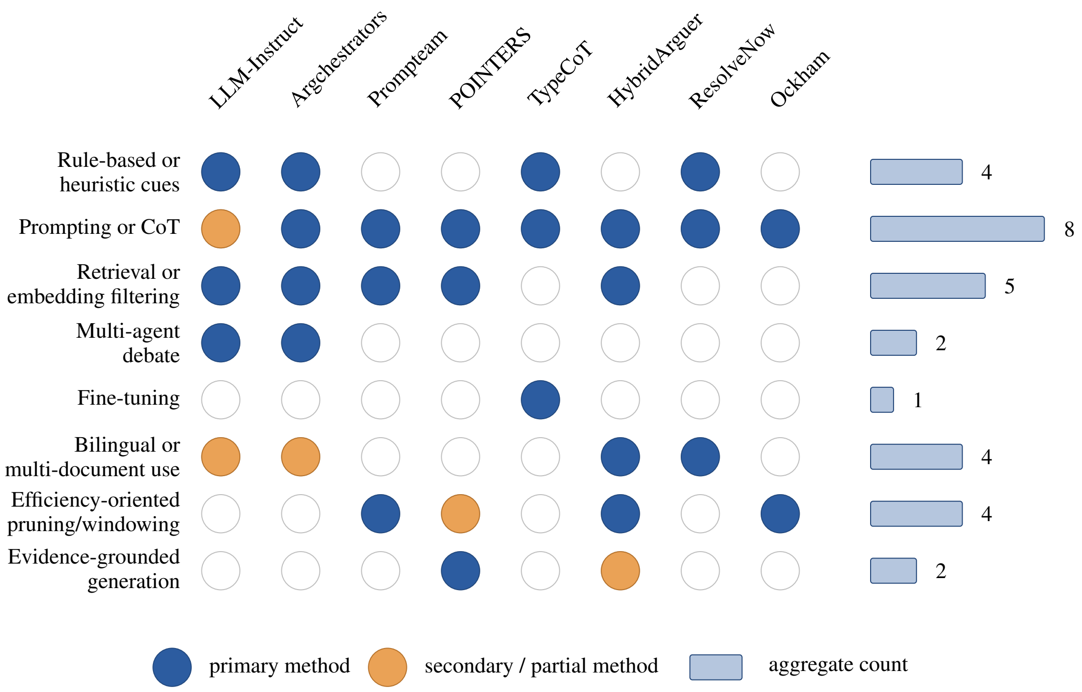


Figure 2: Overview of methodological techniques used by participating teams.

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 ≈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 dimensions, 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 (≈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.

LLM-INSTRUCT at UZH Shared Task 2026: Constraint-Aware Retrieval and Selective Debate for Paragraph-Level Argument Mining

Phuong Huu Vu Tran[1,*,**], **Long Minh Vo**[2,*,**], **Son Nguyen Minh Le**[1], **Hoang Van**[2,3]
[1]Vietnamese-German University, Vietnam
[2]RMIT University Vietnam, Vietnam
[3]VANGIA INNOVATIONS, Vietnam
[*]These authors contributed equally.
[**]Corresponding authors: 10425032@student.vgu.edu.vn, s4215945@rmit.edu.vn

Abstract

We present LLM-INSTRUCT, the winning system for the UZH Shared Task at ArgMining 2026 on paragraph-level argument mining in UN and UNESCO resolutions. The task requires paragraph-type classification, prediction of a subset of 141 official tags, and directed relation prediction under a strict JSON schema setting using only open-weight models up to 8B parameters. We frame the task as constrained structured prediction. The system first narrows the candidate tag space with metadata-aware dense retrieval, then applies constrained decoding with per-dimension caps, escalates only uncertain cases to a three-agent debate branch, and finally validates the output schema. On the official leaderboard, LLM-INSTRUCT ranked **1st overall**, with **1st in F1** and **5th in LLM-as-a-Judge**. During development, our configuration search further improved Task 1b Micro-F1 from 35.83% to 40.08% while keeping the internal Task 2 score at 4.421. The main lesson is simple: reducing the decision space before generation improves both accuracy and submission robustness. Our code and supporting scripts are publicly available at: https://github.com/LLM-Instruct-at-UZH-Shared-Task-2026/Method

1 Introduction

The UZH Shared Task at ArgMining 2026 asks participants to reconstruct argumentative structure in highly formal institutional texts. For each paragraph, a system must determine whether it belongs to the preambular or operative part of a document, assign a subset of 141 predefined thematic tags, and recover directed argumentative relations to other paragraphs in the same document. The task is difficult because the texts are long, the label inventory is closed and structured, and the submission format is strict: non-conforming JSON is not evaluated. In other words, the benchmark requires both semantic plausibility and schema-valid output.

This benchmark is better viewed as a constrained prediction task than as open-ended text generation. The model must reason over long institutional paragraphs, but it must also stay within a fixed label inventory, preserve paragraph indices, and produce schema-valid predictions. We therefore narrow the admissible output space before final generation. Candidate retrieval narrows the tag space; organizer-provided metadata is reused in retrieval and in per-dimension caps; decoding is projected back to the retrieved set; and final validation enforces the submission schema.

This paper makes three contributions. First, it describes the end-to-end pipeline of the first-ranked LLM-INSTRUCT submission under the official shared-task constraints. Second, it identifies the key design choice behind the result: constraining admissible tags before generation instead of asking the model to search the full 141-tag inventory. Third, it reports the development trajectory that exposed the main failure mode of early runs, namely cross-dimension over-prediction.

2 Related Work

Our system combines ideas from constrained decoding, dense retrieval, debate-style reasoning, and argument-mining pipelines. In particular, it is closest to settings that restrict admissible outputs before or during decoding (Geng et al., 2023; Liu et al., 2022), while also drawing on dense retrieval (Karpukhin et al., 2020), debate-style control (Du et al., 2024), and prior argument-mining work (Lawrence and Reed, 2020; Stab and Gurevych, 2017). The difference in this shared task is that semantic plausibility alone is not enough: predictions must also satisfy a strict output schema. For broader context, our pipeline also contrasts with end-to-end and text-to-text approaches to argument mining, including structured prediction as generation (Paolini et al., 2021), end-to-end universal argument mining (Cao, 2023), fine-tuned LLM pipelines for AM (Cabessa et al., 2025), and recent LLM work on relation-based argument mining (Gorur et al., 2025).

3 Proposed Method

Figure 1 summarizes the pipeline. The final leaderboard run has three prediction stages followed by a submission-safety layer. We design the pipeline to satisfy these strict task constraints while keeping generation tightly controlled. We describe its components next.

3.1 Type stage

Each paragraph is first classified as *preambular* or *operative*. In development, we found that a deterministic

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 119–124
July 3, 2026 ©2026 Association for Computational Linguistics

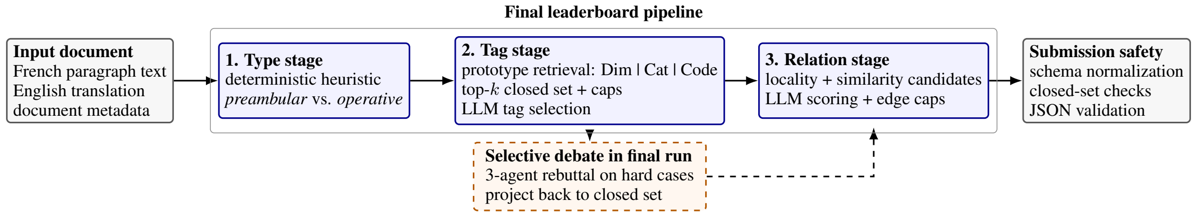


Figure 1: System pipeline. The winning configuration reduces the decision space before final generation: dense retrieval creates an admissible tag set, organizer-provided metadata is encoded in the tag prototypes and reused for per-dimension caps, and high-uncertainty cases are escalated to a three-agent debate branch whose output is projected back onto the same admissible label set before final schema validation.

Table 1: Representative triggers used by the type heuristic.

Rule type	Example opener	Predicted type
French preambular cue	*Considérant ..., Reconnaissant ..., Rappelant ...*	preambular
French operative cue	*Demande ..., Souligne ..., Attire ..., Proclame ...*	operative
English compound cue	*Calls upon ..., Takes note ...*	operative
Numbering pattern	`1.`, `2)`, `(3)` at paragraph start	operative
Fallback	no cue matched	preambular

heuristic was more stable than a pure LLM-based alternative, so the final configuration keeps this step rule-based. Specifically, we classify paragraphs as operative when they match explicit numbered clause patterns or operative cue phrases, and otherwise default to preambular. Table 1 gives representative triggers used by the heuristic. This reduces variance early in the pipeline and avoids propagating unstable type decisions to later stages.

3.2 Metadata-aware tag retrieval and decoding

The tag stage is the core of the system. We read the official tag CSV and keep rows whose CODE field is neither empty nor NA. For each remaining tag t, let d_t, c_t, and y_t denote its released dimension, category, and CODE fields. We textualize the tag prototype as

$$p_t = d_t \,\|\, c_t \,\|\, y_t,$$

where $\|$ denotes string concatenation. We then embed both p_t and the paragraph text with intfloat/e5-base-v2 (Wang et al., 2022), and cosine similarity is then used for dense top-k retrieval (Wang et al., 2022; Karpukhin et al., 2020).

This design makes the retrieval process metadata-aware: dimension and category information is present inside the prototype text before final tag generation. The LLM never predicts over the full 141-tag inventory. Instead, it selects from the retrieved closed set. In the final run, we apply a global cap of 5 tags per paragraph and a per-dimension cap of 2 tags. These fixed controls limit redundancy within a single semantic region and suppress cross-dimension over-prediction. Finally, any decoded label outside the retrieved candidate set is rejected. Together, these controls address the main failure mode observed during development, namely cross-dimension over-prediction.

In addition to retrieving candidate tags, the final tag prompt includes up to three retrieved in-context examples from the training release when their embedding similarity exceeds 0.70. These examples are used only as contextual evidence for how similar paragraphs are tagged; they do not expand the official tag inventory, and the final prediction is still projected back to the retrieved closed set.

3.3 Selective debate branch

We also implement a hard-case route with three open-weight 8B agents-Qwen3-8B (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Ministral-8B-Instruct (Mistral AI, 2024)-inspired by debate-style reasoning (Du et al., 2024).

Under this setup, a debate is triggered only for uncertain paragraphs, namely cases where the top-1 tag margin is low under a development-tuned routing rule, the heuristic and generator disagree on type, or the retrieved candidates show strong overlap across competing tag dimensions. Each agent proposes a type/tag hypothesis, the agents exchange one focused rebuttal round, and a constrained selector accepts only labels that remain inside the retrieved candidate set while reapplying the same per-dimension caps. Put differently, debate never expands the admissible label set; it only helps resolve hard cases within the same retrieved closed set.

3.4 Relation stage

Relation prediction starts from sparse candidate generation rather than exhaustive all-pairs scoring. For clarity, Table 2 summarizes the four official relation labels used in Subtask 2. For each source paragraph, we form a candidate target set by taking all targets within a locality window of one paragraph and adding the top six embedding-similar targets in the same prediction instance. This keeps nearby discourse links available while still allowing non-adjacent semantic links.

The Qwen3-8B generator then scores each candidate pair using only the four official labels in Table 2 or a no-edge option. We keep pairs whose confidence is at least 0.40 and cap each source paragraph at five outgoing edges, ranked by confidence. These filters

Table 2: Official relation labels in Subtask 2.

Label	Informal meaning
supporting	strengthens or justifies another paragraph
complemental	adds compatible information without conflict
modifying	narrows, qualifies, or conditions another paragraph
contradictive	conflicts with or contradicts another paragraph

control the density of the relation graph and prevent relation prediction from becoming an unbounded all-pairs generation problem. Because the held-out test release does not expose gold relation labels, we report relation-output statistics rather than gold precision or recall for this stage.

3.5 Submission safety

The final stage focuses on submission validity. It normalizes schema variants at ingestion, validates required keys before export, enforces the official relation label set, checks paragraph-index consistency, and attempts to repair malformed JSON up to three times when necessary. This component is essential because submissions that violate the required format do not receive an official score.

4 Experiments

4.1 Data and Official Setting

The shared task contains two subtasks. Subtask 1 predicts paragraph type (preambular or operative) and a multi-label subset of 141 official tags. Subtask 2 predicts paragraph-to-paragraph links and labels each directed relation as contradictive, supporting, complemental, or modifying. Official ranking averages an automated F1 metric and an LLM-as-a-Judge score. Because the public leaderboard reports ranks rather than absolute values, we report official rank positions and complement them with internal development metrics.

The organizers release a large *unlabelled* training set drawn from $2,695$ UN resolutions from the UN-RES corpus associated with the SpiritRAG resource (Gao et al., 2025), together with a held-out UNESCO evaluation split. We use the task release as provided by the organizers and do not introduce additional training labels. The texts are provided in French, with English translations provided by the task organizers to support non-French-speaking participants. In our final configuration, Task 1 and Task 2 read the English field whenever it is available and otherwise fall back to the original French paragraph. We prioritize English when available because the organizer-provided translations simplified prompt design, qualitative inspection, and debugging for a non-French-speaking team. The task page also releases a CSV file, evaluation_dimensions_updated.csv, containing the official tag inventory together with dimension and category metadata. Our method uses this file directly in both retrieval and decoding. In the current analyzed artifact, all evaluated instances had English available, so we do not present a separate French-only case study in this version.

4.2 Models and Evaluation Protocol

The final system complies with the task policy of using only open-weight models with at most 8B parameters. The main generator is Qwen3-8B in 4-bit inference mode (Yang et al., 2025). Dense retrieval uses intfloat/e5-base-v2 (Wang et al., 2022). Table 3 reports the deterministic settings of the best performing internal configuration.

We report two complementary views of performance. First, we present the *official leaderboard results* on the test set. Because the organizers have not yet released the exact official scores, we report rank positions rather than absolute scores. Second, we report the *internal development trajectory* that guided system selection and revealed the main failure mode. Following the shared-task setup, our internal evaluation uses Task 1 scores together with a Task 2 score derived from judge-based relation assessment. Code, prompts, and supporting scripts are publicly available at our GitHub repository.

Component	**Setting**
Generator	Qwen3-8B (4-bit)
Reasoning budget	256
Embeddings	intfloat/e5-base-v2
Language path	English if available, else French
Type mode	heuristic
Tag retrieval	cosine top-k, $k = 40$
Tag textualization	Dimensions | Categories | CODE
Tag decoding	threshold 0.33; max 5 tags
Per-dimension cap	max 2 tags per dimension
Relation candidates	window $= 1$, $k = 6$
Relation filtering	threshold 0.40; max 5 edges/source
RAG examples	$k = 3$, minimum cosine score 0.70
Debate	enabled for hard cases
JSON repair	max 3 retries

Table 3: Key settings of the strongest internal configuration and the final leaderboard submission.

4.3 Results

4.3.1 Final Rank on the Official Leaderboard

Table 5 gives the official leaderboard ranks. The main empirical outcome is straightforward: LLM-INSTRUCT ranked **1st overall**. The table also shows the strongest competing teams and our earlier submission, *LLM-Instruct-2*, which corresponds to the earlier Phase2-style configuration.

This pattern is consistent with the design priorities of the system. The gains appear strongest when the benchmark rewards valid structured output. The official rank-1 outcome therefore supports our central claim that schema-aware control layers are useful for paragraph-level argument mining under hard output constraints. It also aligns with the internal trajectory in Table 4, where the earlier submission *LLM-Instruct-2* ranked below the final LLM-INSTRUCT system.

Run	Dominant setting	T1a Acc	T1a F1	T1b P	T1b R	T1b F1	T2 Judge
Phase_0	*Initial baseline from the internal script*	72.19	69.69	**54.92**	26.59	35.83	**4.421**
Phase_1	*Recall-oriented tagging; looser selection increased over-prediction*	85.81	83.21	21.57	30.11	25.13	4.388
Phase_2	*Further recall-oriented tuning; cross-dimension false positives remained high*	85.77	83.16	21.53	30.09	25.10	4.364
Phase_3	*Final constraint-aware run with metadata-aware retrieval, selective debate, per-dimension caps, and closed-set validation*	**86.08**	**83.49**	49.94	**33.47**	**40.08**	**4.421**

Table 4: Internal development trajectory. T1a is paragraph-type classification; T1b is tag assignment. T2 Judge is the internal LLM-as-a-Judge weighted relation score. The main correction from Phase 1/2 to Phase 3 is precision recovery under a similar recall regime, consistent with the over-prediction diagnosis.

Team	F1	Judge	Final
LLM-Instruct	**1**	**5**	**1**
Prompteam	5	1	2
Argchestrators	2	6	3
HybridArguer	4	3	3
LLM-Instruct-2*	7	4	6

Table 5: Official leaderboard ranks from the UZH Shared Task. *LLM-INSTRUCT-2 is our first submission and corresponds to the earlier Phase2-style run.

4.3.2 Development trajectory

Table 4 summarizes the main development phases. We explain the phases explicitly because the most useful lesson from development is not merely that the final run scored higher, but *why* it scored higher.

The pattern is informative. Early recall-oriented configurations raised coarse paragraph-type scores but harmed Task 1b badly. The reason was over-prediction: loose tag selection created many cross-dimension false positives. The final run corrected this failure mode by combining metadata-aware retrieval before generation, retrieved examples in the tag prompt, per-dimension caps during selection, and strict closed-set validation after decoding.

4.3.3 Component diagnostics

We next report a compact component diagnosis, keeping only the results that show clear accuracy or robustness effects. The subset ablations use the same stratified 12-document subset and the same fast decoding setting, so they should be read as relative component evidence rather than absolute final-system scores. Details are in Table 6.

The strongest component effect comes from metadata-aware prototypes: replacing the *Dimension* | *Category* | *CODE* prototype with CODE-only text reduces subset Task 1b F1 by 7.74 points. Retrieved examples are also useful: removing them lowers recall from 22.76% to 16.68% and F1 by 5.17 points. By contrast, closed-set filtering does not materially change subset F1, but it prevents out-of-candidate and invalid raw tags from reaching the final JSON. A full-corpus per-dimension-cap ablation shows a smaller regularization effect: removing the cap changes F1 from 40.26% to 39.84% and increases false positives from 3,529 to 3,575.

Box 4.3.3 shows a representative output.

Variant	P	R	F1	Δ
Subset baseline	37.24	22.76	28.26	–
No RAG examples	37.52	16.68	23.09	-5.17
CODE-only prototype	27.89	16.23	20.52	-7.74
No closed-set filter	37.45	23.08	28.56	+0.30

Table 6: Subset component diagnostics for Task 1b on a stratified 12-document subset under the same fast decoding setting. Removing retrieved examples mainly reduces recall, while replacing metadata-aware prototypes with CODE-only prototypes substantially reduces both precision and F1. Closed-set filtering has little effect on F1 but is retained for schema safety.

Box 4.3.3: Representative model output

Paragraph (English). "While acknowledging that the number of compulsory school years may vary between countries, [the Conference] considers it desirable that the number of actual years of schooling should in no case be less than seven, and notes that this minimum is already exceeded in many countries."
Type. operative
Tags. LAW_REG, ISC_1
Outgoing relations. 2, 4, 7 → complemental.
Interpretation. A legal-regulatory recommendation aligned with nearby policy paragraphs.

We also inspected Phase 3 errors by tag dimension. The largest false-positive counts came from broad dimensions such as Policy theme (623 FP), Teachers (547 FP), Legal frameworks (388 FP), Curriculum (325 FP), and Education level (321 FP). The most frequent false-positive tags were LAW_REG (194 FP), POL_EIE (161 FP), T_OTHER (137 FP), POL_CUR (110 FP), and PEDAG_OTHER (93 FP). This supports the over-prediction diagnosis: broad policy-related dimensions are semantically close to many paragraphs and are therefore easy to over-select under recall-oriented prompting.

Performance also varied strongly by tag frequency. For tags appearing more than 20 times in the internal reference, Phase 3 achieved 57.22% precision, 33.71% recall, and 42.43% F1. For medium-frequency tags with 6–20 references, F1 dropped to 16.61%; for rare tags with at most five references, F1 was only 2.51%. Thus, the constraint-aware design reduced broad over-prediction but did not solve sparse-label recognition.

More concretely, disabling closed-set filtering introduced 8 official tags outside the retrieved candidate set and 5 invalid raw tags in the trace, even though the subset F1 changed only marginally. We therefore retain this step as a reliability guard rather than an accuracy-driven component.

4.4 Output statistics

Operationally, the final submission JSON is organized by recommendation-level TEXT_ID values. The final artifact contains 89 prediction instances spanning 44 unique source documents, where the source document title is read from METADATA.structure.doc_title. All runtime and output counts reported below are computed at this prediction-instance level.

Because the held-out test release does not expose gold relation labels, we cannot compute candidate recall or per-label precision/recall against gold for the final submission. We therefore report descriptive output statistics directly from the final submission artifact. Across these prediction instances, the final system produced 13,323 directed edges among 132,840 possible within-instance paragraph pairs, for a graph density of 10.03%. All four official relation labels are present: 8,949 *complemental*, 4,199 *supporting*, 160 *modifying*, and 15 *contradictive*. Non-adjacent links were common: 9,566 edges, or 71.80%, spanned more than one paragraph.

4.5 Compute report

The final run was executed on 2×NVIDIA GeForce RTX 3090 24GB GPUs and processed the 89 prediction instances in 1:42:03, averaging 68.81 seconds per prediction instance, 2.07 seconds per paragraph, or 34.5 minutes per 1k paragraphs.

5 Discussion

5.1 System Strengths and Advancements

Four implementation choices appear central. First, metadata-aware tag prototypes are important: the CODE-only diagnostic produced the largest observed subset drop. Second, retrieved examples improve recall by supplying paragraph-level usage context without expanding the label inventory. Third, admissibility rules improve reliability by projecting predictions back to the retrieved official candidate set. Fourth, per-dimension caps act as a modest regularizer, reducing false positives on the full corpus even though their absolute F1 effect is smaller than the retrieval-related components.

5.2 Limitations

There are three main limitations in this work. **First**, the relation stage is less mature than the tag stage and remains sensitive to candidate generation and confidence thresholds. **Second**, our component diagnostics are not a full factorial ablation. Several toggles are evaluated on a stratified subset under a faster decoding setting, so they should be interpreted as relative component evidence rather than absolute final-system scores. **Third**, the default path prioritizes English translations when available, so additional cross-lingual analysis on French-only cases would be valuable in future work.

6 Conclusion

We presented a compact, constraint-aware, and submission-safe system for paragraph-level argument mining in UN and UNESCO resolutions. The system combines metadata-aware dense retrieval, constraint-aware decoding, selective debate for high-uncertainty cases, sparse relation prediction, and explicit schema validation. Taken together, these results suggest that under hard output constraints, carefully designed control layers and selectively applied multi-agent reasoning can be as important as stronger generation.

Ethics Statement. Our system is intended for research benchmarking on institutional text, not autonomous legal or policy decision-making. Its design keeps intermediate decisions auditable and makes its constraints explicit.

References

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Lang Cao. 2023. Autoam: An end-to-end neural model for automatic and universal argument mining. In *Advanced Data Mining and Applications*, pages 517–531, Cham. Springer Nature Switzerland.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can large language models perform relation-based argument mining? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534, Abu Dhabi, UAE. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mistral AI. 2024. Model card for ministral-8b-instruct-2410. Accessed 2026-04-16.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *Preprint*, arXiv:2101.05779.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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.

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
July 3, 2026 ©2026 Association for Computational Linguistics

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.

Argchestrators at UZH Shared Task 2026: Efficient Argument Mining in UN Resolutions: A Sub-8B Pipeline using Agentic Debate and Heuristic Retrieval

Bogdan Octavian Grecu[1], **Gerrit Quaremba**[1], **Elizabeth Black**[1], **Denny Vrandečić**[2], **Elena Simperl**[1], **Oana Cocarascu**[1]

[1]Department of Informatics, King's College London
[2]Wikimedia Foundation
Team: Argchestrators
{bogdan.grecu, gerrit.quaremba, elizabeth.black, elena.simperl, oana.cocarascu}@kcl.ac.uk
denny@wikimedia.org

Abstract

The highly formal and negotiated language of United Nations (UN) resolutions presents unique challenges for argument mining. This paper describes our system submitted to the ArgMining 2026 Shared Task: Reconstructing the Reasoning in United Nations Resolutions. Adhering to the strict constraint of utilising open-weight models with $\leq$ 8 billion parameters, we propose a hybrid, compute-efficient architecture powered by Qwen3-8B. For the preambular-operative classification, we implement a set of deterministic rules related to the specificity of UN documents, supplemented by zero-shot prompting to handle edge-case paragraphs that fall outside these heuristics. For tagging, we implement a hierarchical multi-agent debate system. For relation prediction, we deploy a two-stage retrieve-and-rank pipeline, introducing an asymmetric distance-decay heuristic to model the backward-referencing nature of UN legal texts. Our approach shows that careful pipeline engineering can allow highly constrained models to perform sophisticated argumentative reasoning.

1 Introduction

Argument mining in the political and legal domains has focused on extracting implicit reasoning and structural dependencies from formal texts (Held and Habernal, 2025; Liepiņa et al., 2025). United Nations (UN) resolutions contain collective reasoning at an international scale, characterized by a specific tone of discourse and a highly regulated structure (i.e. preambles, negotiated implicit premises, and operative conclusions). The ArgMining 2026 Shared Task focuses on reconstructing these underlying argumentative structures from a corpus of UN resolutions and is split into two subtasks: 1) Argumentative Paragraph Classification: Formulated as a joint classification problem, systems must first perform a binary classification to determine a paragraph's structural role (*preambular* vs. *operative*), followed by a highly granular multi-label classification task to assign relevant thematic dimensions from a predefined 141-tag taxonomy; and 2) Argumentative Relation Prediction: Formulated as a directed graph extraction task, systems must identify logical links between paragraphs and classify the nature of each edge into one of four argumentative relations: *contradictive*, *supporting*, *complemental*, or *modifying*.

The ArgMining 2026 Shared Task requires systems to rely exclusively on open-weight models of up to 8B parameters. Although large-scale proprietary models exhibit robust zero-shot extraction capabilities on intricate schemas, models $\leq$ 8B often suffer from performance degradation when navigating extensive multi-label taxonomies or handling the quadratic computational complexity inherent in pairwise relation extraction (Wei et al., 2022a). Performance can be improved by asking multiple agents, each powered by the 8B model, to reason and debate about more difficult classification tasks, before making a final decision (Du et al., 2023).

To overcome these limitations, we propose a hybrid architecture that heavily filters inputs using deterministic rules and lightweight embeddings, while assigning the LLM to handle high-ambiguity cases. Our system[1] consists of:

1. A **Deterministic-to-Generative** pipeline for preambular and operative classification which uses domain-specific structural heuristics to minimise computational overhead, sending only the cases that do not match these heuristics to a zero-shot LLM fallback.

2. A **Hierarchical Multi-Agent Debate** frame-

[1]Our code is available at https://github.com/grecu-bogdan-13/UN_resolutions_shared_task

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 129–135
July 3, 2026 ©2026 Association for Computational Linguistics

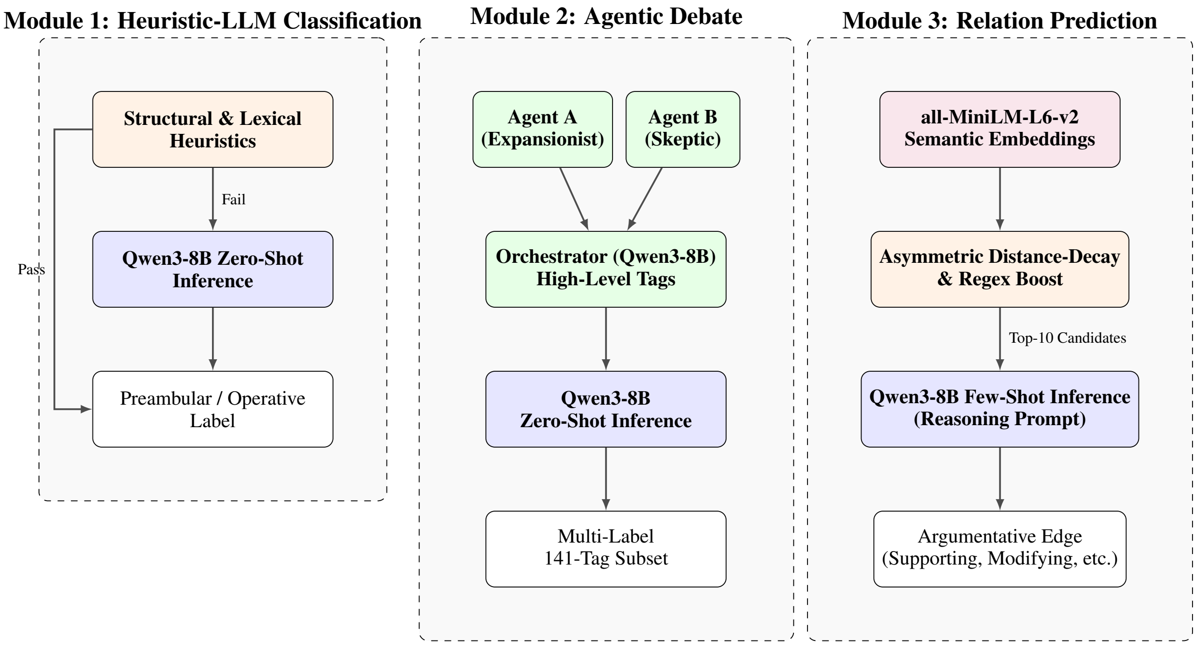


Figure 1: The modular architecture of our hybrid argument mining pipeline. To comply with the strict parameter constraints, we utilize `Qwen3-8B` agents (green) and prompt `Qwen3-8B` (purple), while structural logic relies on deterministic heuristics (orange) and lightweight embedders (pink).

work that asks an 8B model to explicitly reason about broad dimensions before deciding on specific sub-categories.

3. An **Asymmetric Distance-Decay** filtering mechanism that mathematically models the backward-referencing patterns of UN texts, drastically reducing the search space for relation prediction.

2 Methodology

Our system architecture (Figure 1) is modular, with three modules tackling the three components of the shared task. Our system utilises `Qwen3-8B`[2] (Bai et al., 2023) across all components in order to comply with the strict $\leq$ 8B parameter constraint. The training and test data provided by the shared task were introduced in Gao et al. (2025).

2.1 Subtask 1: Paragraph Classification

The classification subtask requires assigning a binary functional role (preambular vs. operative) and selecting from a high-dimensional (141-tags) multi-label taxonomy.

2.1.1 Preambular vs. Operative Classification

As UN resolutions are drafted using highly regulated templates, deploying an LLM for every paragraph is computationally wasteful. For the preambular/operative binary classification, we implement a sequential filtering mechanism. Paragraphs are evaluated against continuation inheritance rules (e.g., clauses beginning with "that" or "que" inherit the context of the preceding paragraph) and strict bilingual lexical triggers (e.g., "acknowledging", "decides"). Paragraphs that satisfy these conditions are classified deterministically. Only the paragraphs that fail these heuristics are passed to Qwen3-8B for zero-shot classification, saving computational resources while ensuring robust handling of formatting anomalies.

2.1.2 Hierarchical Multi-Agent Debate

To navigate the complex 141-tag label space, we deploy a Generator-Discriminator agentic debate architecture. The system decomposes the taxonomy hierarchically. First, two instantiated agents debate the applicability of high-level dimensions:

- **Agent A (Expansionist):** Prompted to maximise recall, proposing any dimension reasonably supported by the text.
- **Agent B (Skeptic):** Prompted to maximise precision, demanding strict justification to include a label.

An *Orchestrator* agent monitors the debate, summarises the reasoning, and decides whether to stop

[2] `https://huggingface.co/Qwen/Qwen3-8B`

or continue the debate. For each selected high-level dimension, we deploy chain-of-thought (Wei et al., 2022b) zero-shot inference to select the specific low-level categories. The prompt includes the high-level dimension, the orchestrator's summary, and detailed instructions for selecting the low-level tags, emphasising precision over recall.

2.2 Subtask 2: Argumentative Relation Prediction

Evaluating every paragraph pair in a document scales quadratically ($O(N^2)$), rendering pairwise LLM evaluation computationally intractable. We solve this using a two-stage retrieve-and-rank pipeline as follows.

2.2.1 Phase 1: Asymmetric Distance-Decay Filtering

We utilise a lightweight sentence transformer `all-MiniLM-L6-v2`[3] to generate initial semantic similarity scores between anchor and candidate paragraphs. However, semantic similarity alone fails to capture the structural flow of UN resolutions, where operative paragraphs frequently refer back to preambular context, but preambular paragraphs rarely reference operative ones. To capture this, we introduce an *Asymmetric Distance-Decay* penalty. In UN texts, paragraphs primarily reference preceding logic. Let d be the difference in index between the anchor and candidate. The decay multiplier $f(d)$ is defined as:

$$f(d) = \begin{cases} e^{-0.05d} & \text{if } d > 0 \\ e^{-0.5|d|} & \text{if } d < 0 \\ 0 & \text{otherwise} \end{cases}$$

This gently decays backward-looking references while harshly penalising forward-looking ones. Explicit regex matches (e.g., "paragraph 5") receive a flat +1.0 score boost. The top-k (k=10 in our experiments) candidates are retained for the next phase.

2.2.2 Phase 2: LLM Inference

The top candidates are dynamically batched and passed to Qwen3-8B. We utilise a reasoning, few-shot prompt that forces the model to generate structured intermediate reasoning before outputting the final JSON classification (*contradictive, supporting, complemental, modifying*, or *none*).

[3] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

3 Experimental Setup

All experiments, including the agent-enabled debate, were performed on an HPC cluster using `Qwen3-8B`. For Subtask 1 debate generation, we utilised a temperature of 0.2 to encourage focused arguments. For Subtask 1 preambular vs. operative classification and for Subtask 2, we utilised a temperature of 0.6 and top-p of 0.95. The prompts used are provided in Appendix A.

4 Results

The task utilises two primary metrics: an F_1 score to evaluate the classification accuracy and relation extraction precision, and an LLM-Judge score to assess the qualitative logical coherence of the generated reasoning. Our system ranked third overall. Table 1 shows the leaderboard results.

Evaluation Metric	Leaderboard Rank
F_1-Score	2
LLM-Judge	6
Aggregate Final Rank	**3**

Table 1: Official evaluation results for our pipeline.

4.1 Detailed performance analysis

This section evaluates the efficacy of our proposed hybrid architecture across the individual subtasks.

4.1.1 Preambular vs Operative Classification

For the binary classification subtask, we deployed a sequential Deterministic-to-Generative pipeline. The system first evaluates whether a paragraph adheres to strict formatting rules indicative of preambular or operative roles (denoted as "Rule-based" in Table 2). Subsequently, it checks for continuation inheritance, identifying clauses that begin with specific trigger words signalling that the paragraph inherits the contextual scope of the preceding text (denoted as "Inheritance"). If a paragraph fails to satisfy these deterministic heuristics, the system utilises a zero-shot fallback, prompting Qwen3-8B to classify the anomaly (denoted as "LLM").

Table 2 details the performance and coverage (the percentage of total paragraphs processed) for each subsystem. The "Deterministic" row aggregates the results of the Rule-based and Inheritance heuristics. We can see that the deterministic heuristics successfully classify the vast majority of the dataset (93.56% coverage) while maintaining exceptionally high accuracy (97.41 F_1). This confirms our hypothesis that the highly regulated, template-driven nature of UN resolutions can be exploited to bypass computationally expensive LLM inference for routine paragraphs. The zero-shot Qwen3-8B fallback exhibits a noticeable performance drop (72.06 F_1) on the remaining 6.44% of the data. This discrepancy suggests that the edge-case paragraphs failing our structural heuristics are inherently ambiguous or irregularly formatted, making them challenging even for the generative model. The hybrid architecture yields a overall F_1 score of 95.48, demonstrating the viability of utilising strict structural filtering to maximise both computational efficiency and extraction accuracy under restricted parameter budgets.

Subsystem	**Performance (Weighted F_1 Score)**	**Coverage (%)**
Rule-based	97.29	85.12
Inheritance	98.72	8.44
Deterministic	97.41	93.56
LLM	72.06	6.44
Overall	95.48	100

Table 2: Weighted F_1 scores and coverage distribution for the preambular/operative binary classification task.

Label	**Weighted F_1 (TN included)**	**Weighted F_1 (TN excluded)**
Education level	94.67	6.53
Education orientation	83.41	0.23
Learning modality	95.78	1.08
Ownership/Provision	96.51	0.17
Teachers	94.47	16.18
Infrastructure & resources	93.85	2.95
Curriculum	87.24	0.49
Pedagogy & assessment	89.96	1.36
Subject domain	96.02	7.09
Cross-cutting themes & skills	91.78	0.81
Policy theme	90.72	4.80
Education system monitoring & evaluation	89.32	0.66
Legal frameworks	95.11	1.43
Stakeholder focus	91.51	8.44
Learner population	95.39	5.54
Total	92.78	3.07

Table 3: Performance on the multi-class classification task by high-level categories, with true negatives (TN) included, and excluded, respectively.

Table 3 presents the evaluation results of the multi-agent debate framework for the 141 predefined tags. Since the number of possible tags is large, and each paragraph typically has only a small number of tags, we will encounter a large number of true negatives. This will artificially increase the F_1 score. In order to provide a more comprehensive evaluation of our system, we will report the F_1 score under two paradigms. The first one removes the true negatives completely and sets them to 0, while the second one takes them into account for our calculation. Overall, our system achieves a weighted F_1 score of 92.78, however, if we ignore the non-assignments of tags, performance drops to 3.07. This indicates that the debate-based architecture tended to over-predict labels.

We also experimented with more neutral "personas" instead of the generator-discriminator roles, but observed considerably more extensive label assignments. Thus, the current framework likely balances the precision-recall trade-off more effectively. Another difficulty was caused by the imbalanced distribution of categories in the dataset. Rare labels with limited support are especially difficult to predict accurately under a zero-shot prompting setup and strict parameter constraints.

Table 3 also details the results across the 15 high-level categories. We find that performance varies substantially across thematic dimensions. The strongest results are obtained for broad and semantically distinctive categories such as *Teachers*, *Stakeholder focus* and *Subject domain*, likely because these categories are comparatively well represented in the corpus and contain strong lexical or policy-oriented cues. Performance is substantially weaker for narrow or low-support categories such as *Education orientation* and *Education system monitoring & evaluation*, suggesting that the hierarchical decomposition strategy is partially successful. By first debating high-level dimensions before selecting low-level tags, the system is able to maintain relatively strong recall despite the large label space (a 141-label multi-label problem). The architecture therefore succeeds at narrowing the search space and guiding the downstream tag prediction process. Nevertheless, the transition from high-level dimensions to specific sub-tags remains noisy, particularly when multiple semantically adjacent categories coexist within the same paragraph.

4.1.2 Argumentative Relation Prediction

Evaluating relation extraction in documents with high paragraph counts presents a unique challenge due to the quadratic ($\mathcal{O}(N^2)$) scaling of possible links. Moreover, the majority of paragraph pairs do not have an argumentative relationship, leading to extreme class imbalance. For a comprehensive evaluation of our system, we report the results using two distinct paradigms. The first focuses exclusively on the model's ability to extract and correctly classify active argumentative edges by ignoring the heavily dominant none class. The second is a Full

Matrix Evaluation, which assesses every possible paragraph combination in the document, factoring in true negative (none-none) agreements.

Relation Class	Precision	Recall	F_1 Score	Support
Complemental	2.40	52.09	4.58	597
Supporting	0.84	35.58	1.64	104
Contradictive	0	0	0	7
Modifying	4.85	3.10	3.78	969
Standard Aggregate Performance (Excluding ‘none’ class)				
Weighted Avg	3.71	22.54	3.92	1677
None	99.42	87.46	93.06	135701
Full Aggregate Performance (Including ‘none’ class)				
Weighted Avg	98.25	86.67	91.97	137378

Table 4: Performance metrics for Subtask 2. The table contrasts the Standard Evaluation (which isolates active relation classes) against the Full Matrix Evaluation (which includes the sparse none class).

The evaluation metrics highlight the severe difficulties associated with highly imbalanced relation extraction under constrained parameter limits. When including the none class, the system achieves a 91.97 weighted F_1 score. However, the Standard Evaluation reveals significant shortcomings in the Phase 2 LLM Inference. The system suffers from critically low precision across all active classes, indicating that while the model captures a moderate portion of valid links, it fundamentally over-predicts active relations, generating a massive volume of false positives. This signals that the Phase 1 Asymmetric Distance-Decay filtering should be more strict, minimising the search space even further, and that the model should be asked to be more "conservative" when considering whether a pair of paragraphs is related or not.

4.2 Performance Analysis and Discussion

The difference between our F_1 rank (2nd) and the LLM-Judge rank (6th) highlights the functional trade-offs inherent in a hybrid, parameter-constrained architecture.

Improving Precision through Heuristics: The second place F_1 rank suggests that utilising deterministic heuristics (such as the lexical triggers and asymmetric distance-decay) effectively constrained the search space. This strategy mitigated the tendency of LLMs to hallucinate argumentative components, a behaviour frequently observed in purely generative solutions.

Limitations in Explanation Quality: The sixth place ranking in the LLM-Judge evaluation suggests a limitation in the system’s capacity for deep qualitative explanation. Although the 8B model successfully performed the classification, relying on it to explicitly reason and explain its choices for the first and last tasks produced justifications that likely fell short of the evaluator’s linguistic and qualitative standards.

Evaluating the Hybrid Approach: These results demonstrate that for highly regulated domains like UN resolutions, hybrid architectures provide superior extraction accuracy (F_1) by utilising the document’s structure to guide the LLM. However, a significant gap remains between efficient, accurate extraction, and the generation of high-quality reasoning under strict parameter limitations.

5 Conclusion

In this paper, we described a hybrid architecture designed to parse the argumentative structure of UN resolutions under strict parameter limitations. By combining deterministic heuristics, hierarchical agentic debate, and asymmetric semantic retrieval, we showed how smaller language models can perform complex, domain-specific reasoning tasks.

Our modular architecture opens several promising avenues for future research across all components of the pipeline. First, the zero-shot LLM fallback could be replaced with either a specialised, distilled model trained exclusively on the edge-cases that fail the deterministic heuristics, or simply by replacing the zero-shot prompt with a few-shot prompt. Second, for the multi-label tagging, incorporating Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) to ground a third “Domain Expert” agent in historical UN databases could help resolve highly technical taxonomy disputes between the Expansionist and Skeptic agents. Furthermore, while our asymmetric distance-decay formula effectively models UN document topology, future work should explore fine-tuning the embedding model via contrastive learning on UN-specific corpora, or replacing the heuristic decay entirely with a learned structural prior using Graph Neural Networks (GNNs). Finally, while our Phase 2 inference focuses on intra-document relations, we could extend our counterfactual reasoning prompts to handle inter-document dependencies, allowing the system to map argumentation graphs across decades of historical UN resolutions.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

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.

Lena Held and Ivan Habernal. 2025. Contemporary LLMs struggle with extracting formal legal arguments. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 292–303, Suzhou, China. Association for Computational Linguistics.

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Rūta Liepiņa, Francesca Galloni, Francesca Lagioia, Marco Lippi, Mariaceleste Musicco, Burcu Sayin, Andrea Passerini, and Giovanni Sartor. 2025. Legal argument mining: Recent trends and open challenges. In *Proceedings of the First Argument Mining and Empirical Legal Research Workshop (AMELR 2025) co-located with the 20th International Conference on Artificial Intelligence and Law (ICAIL 2025)*, volume 4089 of *CEUR Workshop Proceedings*, pages 1–14. CEUR-WS.org.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

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

A Prompt Templates

Below are the prompt templates utilized in our pipeline. Variables provided at runtime (such as the paragraph text or agent names) are denoted by bracketed placeholders (e.g. `{paragraph}`).

Preambular vs Operative LLM Inference

Classify the UN resolution paragraph as 'preambular' or 'operative'.
Provide reasoning and then the final JSON with the key "type".
Paragraph: {paragraph}

High-Level Debate: Agent A (Expansionist)

Debate role: Expansionist. Bias toward broader coverage and include multiple labels when reasonably supported by the UN resolution paragraph.
Turn: {turn}/4
Paragraph: {paragraph}
High-level labels: {$label_1$}, {$label_2$} ...
Current selection (may be empty): {current_selection}
Orchestrator summary so far: {summary}
Modify the current selection by adding and/or removing high-level labels so it best matches the paragraph. Output JSON array of label strings only.

High-Level Debate: Agent B (Skeptic)

Role: Skeptic. Bias toward parsimony and keep labels as few as possible, including a label only when clearly justified by the UN resolution paragraph.
Turn: {turn}/4
Paragraph: {paragraph}
High-level labels: {$label_1$}, {$label_2$} ...
Current selection (may be empty): {current_selection}
Orchestrator summary so far: {summary}
Modify the current selection by adding and/or removing high-level labels so it best matches the paragraph. Output JSON array of label strings only.

High-Level Debate: Agent A (Orchestrator)

ou are the orchestrator for a debate over labels for UN resolution paragraphs. You summarize the debate and decide whether more debate is needed. Return ONLY a JSON object with keys: summary (string), continue (boolean), selection (array). No extra text.

Tag LLM Inference

You are an expert education labeller selecting sub-level labels for ONE high-level label from UN resolution paragraphs. You must be concise, evidence-based, and context-aware of UN resolution language (normative statements, policy commitments, rights framing, implementation language). Follow these rules: 1) Select only from the allowed categories. 2) Use the paragraph as primary evidence and the orchestrator summary as supporting context. 3) Prefer precision over over-labeling, but include multiple categories when clearly supported. 4) Do not invent categories or rely on external facts. 5) Keep reasoning concise and tied to concrete phrases/themes in the paragraph. Return ONLY a JSON object with keys: selection (array of category strings), thinking (string). No extra text.

Relation LLM Inference (Phase 2)

You are mapping the internal logic of a single, isolated UN document.
You will be given an Anchor Paragraph and a Candidate Paragraph.
Determine the relationship between the Candidate and the Anchor.
Output ONLY a JSON object in this exact format: {"relation": "label"}.
The label must be one of: 'contradictive', 'supporting', 'complemental', 'modifying', or 'none'.
SUPPORTING: Candidate provides justification, evidence, or context making Anchor's directive valid.
COMPLEMENTAL: Candidate addresses the same theme as Anchor and adds additional info, without depending on each other.
MODIFYING: Candidate changes, qualifies, restricts, or expands the scope of the Anchor.
CONTRADICTIVE: Candidate asserts something conflicting with the Anchor.
NONE: The paragraphs are not related.
Counterfactual test: Does the existence of the Candidate amend, restrict, or expand the specific mandate established in the Anchor?
- If YES -> 'modifying'.
- If NO (it adds a related but separate action) -> 'complemental'.
Provide your thinking process in <think>...</think> tags, then output the final JSON.
If the value is none, no need to output it.
Here are some examples of paragraphs and their relation
Example 1
Paragraph A: Considering that a certain number of students admitted to secondary schools are not in a position to benefit effectively from the instruction provided therein;
Paragraph B: Deems it necessary, in order to avoid as much as possible errors in orientation and the discouragement that may result, to organize student guidance during the final regulated year of primary education, with the collaboration of the teacher, the physician, and the vocational guidance service, with the decision remaining the responsibility of the family.
Relation: supporting
Example 2
Paragraph A: Considers desirable greater coordination between primary education and secondary education in order to facilitate, especially during the initial years of study, the easy transition from one category of education to another.
Paragraph B: Deems it necessary, in order to avoid as much as possible errors in orientation and the discouragement that may result, to organize student guidance during the final regulated year of primary education, with the collaboration of the teacher, the physician, and the vocational guidance service, with the decision remaining the responsibility of the family.
Relation: complemental
Example 3
Paragraph A: Considers it desirable to improve the selection methods for admission to secondary schools proper. For this selection, the following elements should be taken into account: a) the primary school leaving certificate, as well as the individual report prepared by the primary school teachers, b) an examination conducted according to scientific methods aimed at identifying not only the knowledge acquired but also the candidate's aptitude to continue their studies.
Paragraph B: Draws the attention of educational authorities to the fact that, since any selection involves forced elimination, any student excluded from the secondary schools proper should be directed towards other studies or practical vocational training corresponding to their aptitudes.
Relation: modifying
Example 4
Paragraph A: Indigenous peoples have the right to self-determination. By virtue of that right they freely determine their political status and freely pursue their economic, social and cultural development.
Paragraph B: Nothing in this Declaration may be construed as authorizing or encouraging any action which would dismember or impair, totally or in part, the territorial integrity or political unity of sovereign and independent States.
Relation: contradictive

Prompteam at UZH Shared Task 2026: RAG-Augmented Classification and Cosine-Filtered Relation Prediction for UN Resolutions

Siddhartha Khandelwal
GLA University
Mathura, UP, India
siddharthakhandelwal9@gmail.com

Jyotsana Bhardwaj
GLA University
Mathura, UP, India
jyotsanab15@gmail.com

Abstract

We describe our system for the UZH ArgMining 2026 Shared Task on reconstructing argumentative structure in UN/UNESCO resolutions. The task requires (1) classifying paragraph types and assigning thematic tags from a 141-label taxonomy, and (2) predicting directed argumentative relations between paragraphs. Our pipeline combines a quantised Qwen2.5-7B-Instruct model with retrieval-augmented generation (RAG) backed by FAISS-indexed dense embeddings for few-shot prompting and tag candidate pre-filtering. For relation prediction, we apply a sliding-window cosine pre-filter that reduces the quadratic pair space to near-linear cost. A parallelisable, fault-tolerant pipeline with atomic checkpointing enabled complete processing of 2,959 paragraphs across three concurrent Kaggle T4 sessions despite 12-hour GPU limits. Our system achieved **2nd place** overall on the shared task leaderboard.

1 Introduction

The UZH ArgMining 2026 Shared Task (Gao and others, 2026) targets the analysis of argumentative structure in UN/UNESCO resolutions. **Subtask 1** requires per-paragraph prediction of (a) a binary structural type (*preambular* vs. *operative*) and (b) a multi-label subset from a 141-label thematic taxonomy spanning 15 dimensions. **Subtask 2** requires identifying directed relations between paragraph pairs from four types: *supporting*, *contradictive*, *complemental*, and *modifying*. The final ranking averages an automated F1 score and an LLM-as-a-Judge score (0–100) that rates chain-of-thought reasoning quality. All teams must use open-weight models with $\leq$8B parameters.

Our contributions are:

- A RAG-based few-shot classification pipeline using FAISS-indexed dense embeddings for tag candidate pre-filtering and dynamic example retrieval (§3.3).
- A cosine-similarity pre-filter reducing relation prediction from $O(N^2)$ to near-linear cost (§3.4).
- A fault-tolerant, parallelisable design with atomic checkpointing across Kaggle's 12-hour GPU limit (§3.5).
- **2nd place** on the shared task leaderboard.

2 Related Work

Argument mining has progressed from essay-level parsing (Stab and Gurevych, 2017) and MST-based discourse prediction (Peldszus and Stede, 2015) to large-scale political text analysis (Lawrence and Reed, 2020). The UN-RES corpus (Gao and others, 2026) extends argument mining to multilingual UN resolutions with fine-grained thematic tagging. RAG pipelines (Lewis et al., 2020) ground LLM predictions in retrieved evidence; dynamically selecting semantically similar demonstrations outperforms random sampling in few-shot settings (Rubin et al., 2022). We apply this paradigm jointly to tag candidate retrieval and few-shot example selection, and reduce the quadratic cost of pairwise relation prediction via embedding-based cosine pre-filtering.

3 System Description

3.1 Data

The shared task provides both a training set and a held-out test set in JSON format (Gao and others, 2026). The **training data** consists of 2,695 parsed UN resolutions as raw text in French, drawn from the UN-RES dataset (Gao et al., 2025), with machine-generated English translations produced using Helsinki-NLP/opus-mt-fr-en. Use of the training data is unrestricted; the organisers encourage techniques with strong LLM reasoning focus

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 136–139
July 3, 2026

(e.g. RAG, in-context learning). The **test data** comprises 45 parsed documents (resolutions and recommendations) from the UNESCO International Bureau of Education's International Conference on Education (1934–2008), each containing up to three resolutions, annotated at paragraph level in French (with English translations generated using gpt-4.1-mini). In total, the test set contains 89 individual resolutions spanning 2,959 paragraphs. Paragraph-level annotations include a binary structural type (*preambular* vs. *operative*), multi-label thematic tags from a 141-label taxonomy across 15 dimensions (provided in a CSV file), and directed inter-paragraph relations from four categories: *supporting*, *contradictive*, *complemental*, and *modifying*.

3.2 Model Stack

Our LLM is `Qwen2.5-7B-Instruct` (Qwen Team, 2025) in INT8 quantisation (Dettmers et al., 2022) (~7 GB VRAM). Sentence embeddings use `all-mpnet-base-v2` (768-d) (Reimers and Gurevych, 2019) with FAISS `IndexFlatIP` (Johnson et al., 2021) for exact cosine search.

3.3 Subtask 1: Classification

Each paragraph passes through a four-stage pipeline (Figure 1): (1) **Embed** the paragraph into a 768-d vector; (2) **Tag retrieval** via cosine search returns the top-20 candidate tags; (3) **RAG few-shot** retrieves the 5 most similar annotated paragraphs as demonstrations; (4) **LLM classify** with a structured prompt yields a JSON with fields `think` (chain-of-thought), `type`, and `tags`. A heuristic fallback (first-word rule) applies if all three JSON-parse retries fail; output tags are sanitised against the taxonomy.

Figure 1: Subtask 1 pipeline per paragraph.

Prompt design. The system prompt defines the two paragraph types with canonical opening keywords, tag selection rules emphasising precision, and a mandatory `think` template with four reasoning steps (Wei et al., 2022): identify keyword → decide type → evaluate candidates → justify tags. Few-shot examples are ordered by descending similarity (Rubin et al., 2022).

3.4 Subtask 2: Relation Prediction

For each paragraph B, we determine which earlier paragraph A shares an argumentative link. Two strategies reduce the $O(N^2)$ cost (Figure 2):

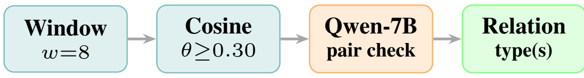


Figure 2: Subtask 2 pipeline per paragraph pair.

Sliding window. Paragraph B is compared only against the preceding w=8 paragraphs, exploiting sequential rhetorical structure. **Cosine pre-filter.** Only pairs with cosine similarity $\geq \theta = 0.30$ reach the LLM, eliminating ~65% of candidates. The prompt requests four-step reasoning: summarise A → summarise B → compare themes → assign relation(s).

3.5 Engineering Design

Parallel chunking. Both tasks distribute work across N=3 concurrent Kaggle notebooks via modulo indexing ($i \bmod N = K$), with a merge step deduplicating by composite key (`doc_id||para_id`). Full hyperparameter settings are provided in Appendix A.

Atomic checkpointing. Results are written to a `.tmp` file and atomically renamed, preventing corruption on session termination. On resume, processed entries are skipped.

4 Results

Our system achieved **2nd place** on the ArgMining 2026 leaderboard.

4.1 Subtask 1: Classification

Metric	Value
Total paragraphs	2,959
Documents	89
Duplicates	0
Preambular / Operative	58.3% / 41.7%
Avg. tags per paragraph	2.4
Avg. `think` length	74 words

Table 1: Subtask 1 coverage and distribution.

Table 1 shows full coverage of all 2,959 paragraphs with zero duplicates after the composite-key fix. The type split (58.3% preambular) is consistent with UN resolution structure. The average of 2.4 tags per paragraph reflects focused multi-label predictions; the 74-word `think` field provides sufficient reasoning depth for LLM-as-a-Judge scoring.

4.2 Subtask 2: Relations

Relation Type	Share
Supporting	44%
Complemental	31%
Modifying	18%
Contradictive	7%
Pairs eliminated by pre-filter	~65%

Table 2: Subtask 2 relation distribution.

Table 2 shows that *supporting* relations dominate (44%), as expected in consensus-driven UN documents. *Contradictive* relations are rare (7%). The cosine pre-filter eliminated ~65% of candidate pairs, substantially reducing LLM inference cost.

4.3 Error Analysis

Two recurring failure modes were identified. First, with only k=20 tag candidates pre-filtered per paragraph, rare labels (e.g., `INFRA_WASH`) may never enter the candidate set. Second, the cosine threshold occasionally filters out argumentatively related pairs with dissimilar vocabulary, particularly for *contradictive* relations.

5 Conclusion

We presented a complete, fault-tolerant pipeline for the UZH ArgMining 2026 Shared Task, achieving 2nd place by combining a quantised 7B instruction-tuned model with FAISS-backed retrieval-augmented generation and cosine-pre-filtered relation prediction. Key takeaways: (i) composite keys are essential for paragraph-level checkpointing across multi-document corpora; (ii) cosine pre-filtering reduces relation prediction cost to near-linear without substantial recall loss; (iii) explicit multi-step `think` prompts produce structurally richer reasoning traces for LLM-as-a-Judge evaluation. Future work includes fine-tuning a small model on the UN-RES training set, exploring cross-document relation links, and learning a calibrated cosine threshold via cross-validation.

Limitations

Tag recall. With 141 tags and only k=20 candidates pre-filtered by embedding similarity, rare thematic tags may be systematically missed. An ensemble of multiple embedding queries or a hierarchical tag-grouping strategy could improve recall.

Relation and reasoning quality. The cosine similarity threshold θ=0.30 was set empirically without cross-validation, which may affect the precision–recall trade-off. Additionally, while prompts enforce structured reasoning, the quality of generated `think` chains remains uneven. Future work could incorporate learned thresholds or reranking models, along with improved reasoning elicitation techniques.

Reproducibility. Our pipeline depends on INT8 quantisation via `bitsandbytes`, which introduces non-deterministic rounding across hardware. As a result, outputs may vary slightly across GPU architectures.

Acknowledgments

We thank the UZH ArgMining 2026 shared task organisers for providing the dataset and evaluation infrastructure. We also thank Kaggle for providing free GPU compute resources that made this work possible.

References

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Yingjia Gao and others. 2026. Reconstructing the reasoning in UN resolutions: A shared task on argument mining. In *Proceedings of the ArgMining Workshop at ACL 2026*. Shared task description paper.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

John Lawrence and Chris Reed. 2020. 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. In *Advances in Neural Information Processing Systems*.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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 (EMNLP)*, pages 3982–3992. Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Appendix

A Hyperparameters

Table A lists the key hyperparameters used in our pipeline.

Component	Parameter	Value
LLM	Model	Qwen2.5-7B
	Quantisation	INT8
	Max new tokens	1,536
	Temperature	0.05
	Top-p	0.9
	Repetition penalty	1.1
	JSON retries	3
Embeddings	Model	mpnet-base-v2
	Dimensions	768
	Batch size	256
Subtask 1	RAG top-k	5
	Tag candidates	20
Subtask 2	Window size w	8
	Cosine threshold θ	0.30

Table A: Key hyperparameters.

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

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

Abstract

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

1 Introduction

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

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

- **Subtask 1: Argumentative Paragraph Classification.** For each paragraph, systems must predict (a) whether it is *preambular* or *operative*, and (b) assign a subset of 141 predefined education-related tags as a multi-label classification problem.
- **Subtask 2: Argumentative Relation Prediction.** Given a paragraph, systems must predict which other paragraphs it is logically related to (identifying them by their paragraph indices), and label each link with one or more argumentative relation types: *contradictive*, *supporting*, *complemental*, or *modifying*.

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

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

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 140–144
July 3, 2026

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

2 Dataset

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

3 System Architecture

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

3.1 Subtask 1: Argumentative Paragraph Classification

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

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

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

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

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

3.2 Subtask 2: Argumentative Relation Prediction

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

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

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

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

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

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

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

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

3.3 Implementation and Infrastructure

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

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

4 Engineering Challenges: Context Limit Recovery

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

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

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

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

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

5 Final Evaluation Results

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

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

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

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

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

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

6 Conclusion

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

7 Limitations and Future Work

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

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

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

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

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

References

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

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

Yingqiang Gao, Fabian Winiger, Patrick Montjourides, Anastassia Shaitarova, Nianlong Gu, Simon Peng-Keller, and Gerold Schneider. 2025. Spiritrag: A q&a system for religion and spirituality in the united nations archive. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 26–41.

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

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

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

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

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

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

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

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

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

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
July 3, 2026

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-*

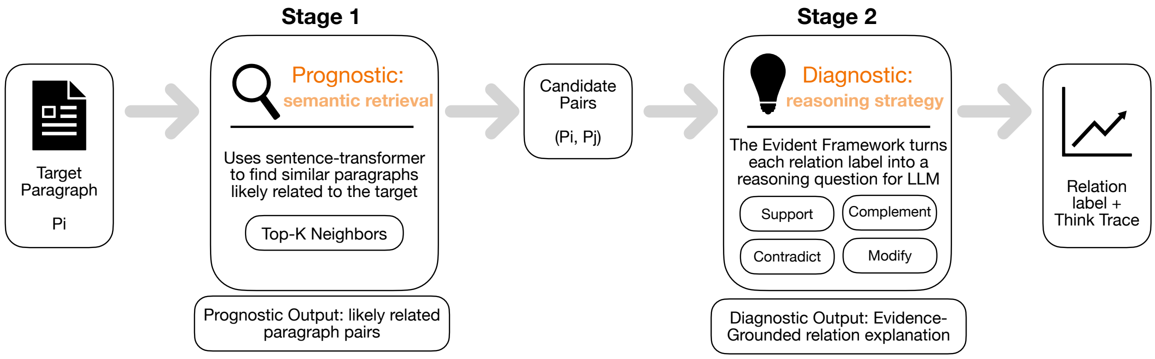


Figure 1: **The two-stage Evident framework** for uncovering argumentative links between resolution paragraphs. Given a target paragraph P_i, the *Prognostic* stage uses a sentence-transformer to retrieve its top-K semantically similar neighbors, producing candidate pairs (P_i, P_j). The *Diagnostic* 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.

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 (≤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
July 3, 2026 ©2026 Association for Computational Linguistics

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 (≤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

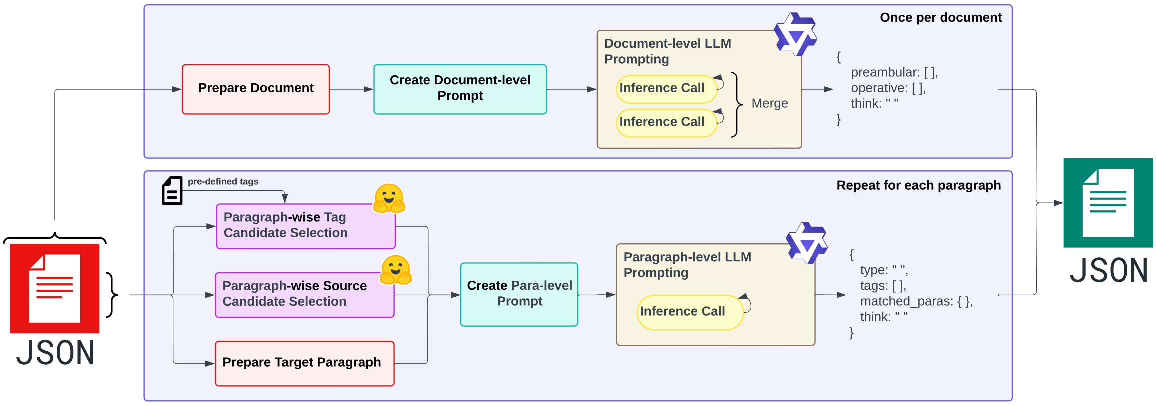


Figure 1: **System Architecture**: The figure illustrates the four-stage end-to-end pipeline that processes a JSON document to produce a structured output. Purple modules correspond to candidate selection stages (Stages 2 and 3), where a multilingual embedding model retrieves top tag candidates and source 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. Acc.	Para. Cls. Stab.	Para. Cls. Reason	Tag Assign. Acc.	Tag Assign. Stab.	Tag Assign. Reason	Rel. Pred. Acc.	Rel. Pred. Stab.	Rel. Pred. 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.

Author Index