

Topological Ordering for Explicit Global Structural Reasoning in Argument Relation Identification

Zhiqiang Shi

Department of Informatics
King's College London
zhiqiang.shi@kcl.ac.uk

Oana Cocarascu

Department of Informatics
King's College London
oana.cocarascu@kcl.ac.uk

Abstract

Argument relation identification (ARI) aims to recover the directed structure of argumentative discourse, yet most approaches, including recent large language models, treat relations as local decisions and rely on implicit reasoning to enforce global coherence. This often leads to structurally inconsistent argument graphs. We reframe ARI as a global structural reasoning problem and introduce topological ordering as an explicit prediction target. Extensive experiments show consistent improvements in ARI performance, edge-level accuracy, and global structural fidelity across models and learning paradigms. In addition, the predicted topological orders provide faithful, verifiable reasoning traces that are internally consistent with model predictions and aligned with gold argument structure, offering an interpretable alternative to free-form chain-of-thought explanations.¹

1 Introduction

Argument relation identification (ARI) is a task in argument mining (AM) that focuses on determining how argumentative components, such as claims and premises, are connected (Stab and Gurevych, 2014; Carstens and Toni, 2015). Accurately identifying argumentative relations is essential for constructing coherent argument graphs that capture the logical flow of discourse, enabling applications in debate technologies, legal analysis, decision support, and explainable AI (Lawrence and Reed, 2019).

ARI is challenging due to its inherently global nature: argumentative relations are not independent, but form a directed graph whose structure must be coherent as a whole. However, traditional neural approaches typically formulate ARI as a set of local or pairwise decisions, determining relations in isolation (Stab and Gurevych, 2014; Park and Cardie, 2018). Even recent approaches using

large language models (LLM) rely on implicit structural reasoning, with global constraints weakly induced or enforced post-hoc (Liu et al., 2023; Gorur et al., 2025; Cabessa et al., 2025). Thus, models often yield argument graphs that are cyclic, directionally inconsistent, or globally incoherent.

We argue that these limitations stem from a mismatch between the structural nature of the task and the learning signals provided. Argument graphs are naturally directed and acyclic, imposing a partial order on argumentative components via reachability: premises precede the claims they support, and conclusions follow the reasoning that underpins them. Yet, this partial-order structure is rarely enforced during training or inference. Instead, models must derive global coherence from local predictions, with limited inductive bias towards structurally valid reasoning.

In this paper, we reframe ARI as a global structural reasoning problem and propose to make global structure explicit by introducing **topological ordering** as a dedicated prediction target. A topological order is a linearization of a directed acyclic graph that satisfies all reachability constraints. Consequently, predicting any valid topological order requires reasoning over the global structure and transitive dependencies of the argument graph, rather than over individual relations alone.

We propose a model-agnostic framework that can be integrated with different learning paradigms, in-context learning (via structural prompting) and supervised fine-tuning (via multitask fine-tuning). This formulation requires no specialized graph architectures or decoder modifications, and exposes structural coherence directly through the prediction target. This design makes the method applicable across LLMs of varying sizes and architectures.

Our evaluation on three benchmark datasets shows that incorporating topological reasoning consistently improves ARI performance. It also increases edge-level accuracy, reduces global struc-

¹Code can be found at: https://github.com/zqshizhiqiang/topological_argument_mining.

tural errors, and yields predictions that are internally coherent and aligned with gold argument structure. Moreover, by generating topological orders alongside predicted relations, our approach provides explicit and faithful reasoning traces that directly reflect the model’s structural decisions. These interpretable reasoning paths allow for fine-grained analysis of common ARI error patterns and offer a transparent alternative to opaque chain-of-thought explanations, grounding model reasoning in the underlying argument graph.

Our **contributions** are as follows: **1)** We propose topological ordering as an explicit structural reasoning objective for ARI, augmenting ARI with the prediction of a globally coherent partial order over argument components; **2)** We present a model-agnostic framework for incorporating global coherence into LLM-based ARI, applicable to both in-context learning and supervised fine-tuning; **3)** We conduct an extensive evaluation showing that explicit topological reasoning improves accuracy, structural fidelity, and explainability in ARI.

2 Related Work

Argument Relation Identification (ARI). Early work used supervised models (conditional random fields (Stab and Gurevych, 2017), SVMs (Stab and Gurevych, 2014)) trained on ARI datasets. Later, research shifted to fine-tuning transformer-based pretrained models and contextualized embeddings (Ruiz-Dolz et al., 2021; Bao et al., 2021, 2022; Liu et al., 2023). Recently, LLMs have enabled ARI via in-context learning (Otiefy and Alhamzeh, 2024; Gorur et al., 2025) and supervised fine-tuning (Cabessa et al., 2025). ARI approaches fall into two main categories. The first treats ARI as a pairwise classification task, processing argument pairs independently to produce relation labels (Stab and Gurevych, 2014; Ruiz-Dolz et al., 2021; Otiefy and Alhamzeh, 2024; Gorur et al., 2025), with global structural constraints enforced post-hoc, e.g. via integer linear programming (Stab and Gurevych, 2017). The second incorporates argument context, but weakly encodes global argument structure, without explicit guidance during learning (Liu et al., 2023; Cabessa et al., 2025). These limitations motivate the need for ARI methods that model the global structure of argumentative discourse rather than relying solely on local predictions.

Graph-grounded LLM Reasoning. Recent works show that LLMs benefit from explicit graph struc-

ture during reasoning. Luo et al. (2024) propose Reasoning on Graphs that prompts models to generate relation paths grounded by knowledge graphs (KG) as faithful reasoning plans, whilst Li et al. (2024) develop a KG-integrated collaboration by integrating graph knowledge directly into multi-step reasoning. Han et al. (2025) propose Reasoning with Graphs to construct explicit context graphs for multi-hop reasoning while Fatemi et al. (2024) demonstrate that LLM performance is highly sensitive to how graph structure is linearized. Besta et al. (2025) propose a taxonomy of structure-guided prompting schemes, showing that graph-structured reasoning improves performance and interpretability compared to linear chain-of-thought (CoT). In these methods, LLMs use graph structure to improve reasoning, but are not required to predict the global graph structure itself. In contrast, we propose topological ordering as an explicit prediction target, moving graph-grounded reasoning from using structure to generating it.

3 Methodology

3.1 Topological Ordering as a Structural Reasoning Objective

Our approach enhances ARI by introducing explicit topological prediction to model the global structure of argumentative discourse. By requiring the model to generate a topological order over the argument relation graph, we expose global dependencies directly during inference and training. In this section, we describe the implementation strategies for in-context learning and supervised fine-tuning.

The core idea is to require LLMs to reason explicitly about both local argumentative relations and global argumentative structure. While ARI focuses on predicting relations between individual components, a topological order over the argument graph captures the overall logical flow induced by these relations. Combining relation prediction with explicit global ordering encourages coherent structural reasoning, improving ARI performance while generating interpretable outputs aligned with the argument’s logical flow.

Motivation for Topological Ordering as a Global Constraint. Argument graphs can be formalized as directed acyclic graphs $G = (V, E)$, where V is the set of argumentative components ($|V| = N$) and $u, v \in V$ denote individual components. A topological order is a mapping $\pi : V \rightarrow \{0, \dots, N-1\}$ such that every gold edge $u \rightarrow v$ satisfies $\pi(u) <$

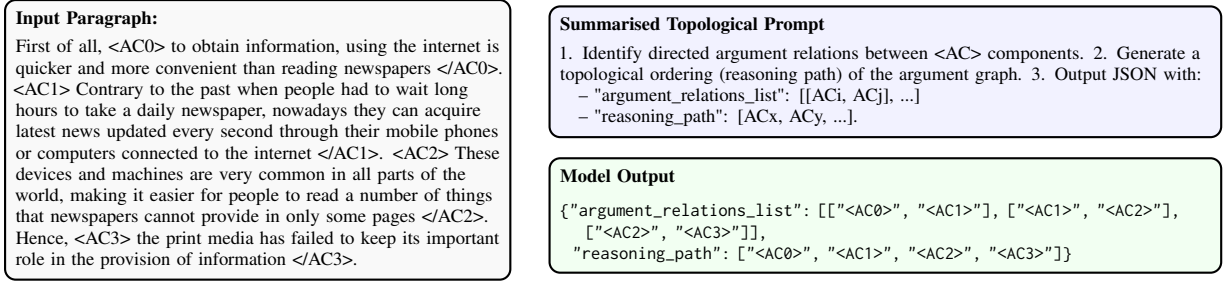


Figure 1: Overview of in-context learning: input paragraph from the PE dataset on the left, and the summarised topological-ordering prompt and model output on the right.

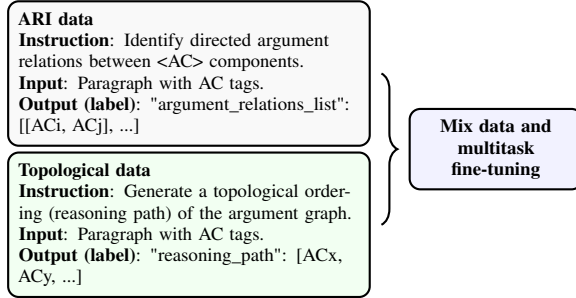


Figure 2: Overview of incorporating topological ordering for supervised fine-tuning.

$\pi(v)$. This single sequence π implicitly encodes $O(N^2)$ pairwise precedence constraints over all component pairs, providing a lightweight but globally coherent structural signal. By construction, any predicted relation must satisfy $\pi(u) < \pi(v)$, preventing cycles and direction reversals and thus enforcing the acyclicity and directionality required of argument graphs. Thus, predicting a topological order acts as a minimal global regularizer, injecting essential structural constraints while allowing the model to freely generate local relations.

In-context learning: We use a single prompt to instruct the model to output the list of argument relations and then generate a topological ordering of the argument graph (see Figure 1 for an example).

Supervised fine-tuning: We incorporate topological ordering as an explicit structural prediction objective during training. This consists of two steps: data generation and multitask training. For each document, we create an additional training instance with the topological order of the gold argument graph as the target. We combine these structural prediction instances with standard ARI examples into a single training set, allowing the model to jointly learn relation identification and global structural reasoning (see Figure 2).

3.2 Structural Evaluation

We report normalized graph edit distance, a standard measure of graph similarity, and introduce two new order-based metrics to evaluate the generated argument graph: edge-order consistency and gold-order agreement.

Notation. Given a document, let V be the set of argumentative components (nodes), $E_{\text{gold}} \subseteq V \times V$ the set of gold directed edges, and $E_{\text{pred}} \subseteq V \times V$ the model’s predicted edges. We denote a total order (linearization) over a subset of nodes predicted by the model as $\pi : V_\pi \rightarrow \{0, 1, \dots, |V_\pi| - 1\}$, where a smaller value $\pi(u)$ indicates that node u appears earlier in the ordering. Let $\rightsquigarrow$ denote reachability in the gold graph, i.e. the transitive closure of E_{gold} . Unless otherwise specified, evaluation is conducted at the document level and aggregated by micro-averaging over the corpus.

Normalized graph edit distance (global structure). To assess structural similarity between the predicted and gold graphs, we compute the normalized graph edit distance (nGED), defined as the minimum number of edge insertions, deletions, or direction flips required to transform the predicted graph into the gold graph, normalized by a size term (e.g. $|E_{\text{gold}}|$):

$$\text{nGED}(E_{\text{pred}}, E_{\text{gold}}) = \frac{\min \# \text{ops}(E_{\text{pred}} \Rightarrow E_{\text{gold}})}{\max\{1, |E_{\text{gold}}|\}}$$

Lower values imply fewer global structural errors.

Edge-order consistency (EO) (internal coherence). Given a model-predicted order π , an edge $u \rightarrow v$ obeys π iff $u, v \in V_\pi$ and $\pi(u) < \pi(v)$. Let $\mathcal{E}_\parallel = \{(u, v) \in E_{\text{pred}} \mid u, v \in V_\pi\}$ and $\mathcal{E}_{\text{obey}} = \{(u, v) \in \mathcal{E}_\parallel \mid \pi(u) < \pi(v)\}$. We report the edge-order consistency rate: $\text{EO}(\pi, E_{\text{pred}}) = \frac{|\mathcal{E}_{\text{obey}}|}{|\mathcal{E}_\parallel|}$, indicating how well the model’s predicted edges align with its predicted global order.

Gold-order agreement (GO) (global reasoning quality). We define the gold partial order by reachability: $u \prec_{\text{gold}} v$ iff $u \rightsquigarrow v$ (i.e. there exists a

Model	AbstRCT	CDCP	PE
Gemma 3 12B AR	0.585	0.549	0.480
Gemma 3 12B topo	0.612	0.559	0.495
Gemma 3 27B pairwise	0.440	0.490	0.351
Gemma 3 27B CoT	0.650	0.550	0.578
Gemma 3 27B AR	0.655	0.542	0.487
Gemma 3 27B topo	0.719	0.565	0.586
LLaMA 3.3 70B pairwise	0.373	0.490	0.314
LLaMA 3.3 70B CoT	0.648	0.532	0.488
LLaMA 3.3 70B AR	0.642	0.533	0.485
LLaMA 3.3 70B topo	0.691	0.549	0.548
Qwen 3 235B AR	0.760	0.561	0.647
Qwen 3 235B topo	0.794	0.598	0.671
GPT 4.1 AR	0.757	0.590	0.618
GPT 4.1 topo	0.780	0.602	0.663

Table 1: Results for in-context learning. Since the pairwise method requires numerous LLM calls and CoT produces long outputs, significantly increasing the cost, we evaluate them only on a subset of models.

directed path from u to v in E_{gold}). We define the set of gold precedence pairs as $\mathcal{P}_{\text{gold}} = \{(u, v) \in V \times V \mid u \rightsquigarrow v\}$. For pairs whose endpoints appear in π , we compute the gold-order agreement as the pairwise precedence accuracy, which measures how well the predicted order aligns with the annotated argumentative flow, as follows: $\text{GO}(\pi) = \frac{|\mathcal{P}_{\text{obey}}|}{|\mathcal{P}_{\parallel}|}$ where $\mathcal{P}_{\parallel} = \{(u, v) \in \mathcal{P}_{\text{gold}} \mid u, v \in V_{\pi}\}$ and $\mathcal{P}_{\text{obey}} = \{(u, v) \in \mathcal{P}_{\parallel} \mid \pi(u) < \pi(v)\}$.

4 Experiment Setup

Datasets We evaluate our approach on three widely used AM datasets from distinct domains: AbstRCT (Mayer et al., 2020), CDCP (Park and Cardie, 2018), and Persuasive Essays (PE) (Stab and Gurevych, 2014). AbstRCT is a biomedical corpus comprising 650 scientific abstracts annotated with argumentative components and relations. Following Liu et al. (2023), we merge MajorClaim and Claim into a single class, and adopt the neo test set as in Morio et al. (2022). CDCP consists of 731 user-generated comments from online policy discussions. PE (Stab and Gurevych, 2014) contains 402 essays with 1,833 paragraphs annotated for argumentative components and relations. We use paragraph-level data without paragraph tags.

Models For in-context learning, we use Gemma 3 (12B and 27B), LLaMA 3.3 (70B), Qwen 3 235B A22B Instruct 2507, and GPT 4.1, accessed via public APIs. All models use greedy decoding, temperature = 0.0, top_p = 1.0. For supervised fine-tuning, we adapt LLaMA 3.1, LLaMA 3.2, and Qwen 2.5 of various sizes using LoRA (Hu et al., 2022) for parameter-efficient training. For larger

models such as LLaMA 3.1 8B and Qwen 2.5 7B, we apply QLoRA (Dettmers et al., 2023) with 4-bit quantization. All experiments are conducted on a single NVIDIA A100 GPU with 40 GB VRAM. For supervised fine-tuning, we report the mean and sample standard deviation across three runs.²

Baselines For in-context learning, we compare our method (topo) against three baselines. We follow Otiefy and Alhamzeh (2024), which frames ARI as a pairwise classification task (pairwise), and prompt the model with all component pairs in the input paragraph to determine whether an argumentative relation exists between them.³ We also include a baseline that uses a similar prompt but requires the LLM to output only argument relations without producing a topological ordering (AR) and a chain-of-thought baseline (Wei et al., 2022) where the LLM generates its reasoning steps before predicting argument relations (CoT).⁴ For supervised fine-tuning, we use Cabessa et al. (2025) as our LLM baseline (SFT) and compare against the strongest non-LLM systems: Bao et al. (2021) on CDCP and Liu et al. (2023) on PE.

5 Results and Discussion

We evaluate the impact of topological ordering on ARI using in-context learning and supervised fine-tuning, analyzing ARI performance and the structural quality of the predicted graphs. For argument graphs, we report edge F_1 (i.e. F_1 score on directed edges) and the structural metrics defined in Section 3.2. For the ARI task, we report F_1 score and macro-average over different classes, following Cabessa et al. (2025).⁵

5.1 Overall performance

5.1.1 ARI performance

In-context learning. Table 1 shows that incorporating topological sorting into the prompt consistently improves ARI performance across all datasets and model sizes. These results indicate that explicitly guiding models to reason about global structure during inference enhances their ability to identify argument relations accurately and consistently.

²Prompting and fine-tuning details are in Appendix C.

³We do not compare against Gorur et al. (2025) as they assume a fixed set of argument pairs and classify relations only among these predefined pairs. In our setting, relations must be inferred among all components within a paragraph.

⁴All prompts are in Appendix A.

⁵See Appendix D for how macro F_1 score is computed.

Model	CDCP	PE
Best non-LLM results	0.678	0.827
LLaMA 3.2 1B SFT	0.595 ± 0.003	0.774 ± 0.006
LLaMA 3.2 1B topo	0.614 ± 0.003	0.779 ± 0.003
LLaMA 3.2 3B SFT	0.672 ± 0.008	0.756 ± 0.003
LLaMA 3.2 3B topo	0.681 ± 0.005	0.793 ± 0.002
LLaMA 3.1 8B SFT	0.714 ± 0.003	0.811 ± 0.007
LLaMA 3.1 8B topo	0.722 ± 0.001	0.824 ± 0.005
Qwen 2.5 1.5B SFT	0.521 ± 0.003	0.780 ± 0.002
Qwen 2.5 1.5B topo	0.585 ± 0.010	0.791 ± 0.006
Qwen 2.5 7B SFT	0.712 ± 0.009	0.798 ± 0.004
Qwen 2.5 7B topo	0.718 ± 0.006	0.814 ± 0.004

Table 2: Supervised fine-tuning results across 3 runs.

Model	CDCP		PE	
	F_1	nGED	F_1	nGED
Gemma3 27B AR	0.146	2.324	0.176	1.267
Gemma3 27B topo	0.192	2.051	0.339	1.001
Qwen3 235B AR	0.185	2.035	0.443	0.809
Qwen3 235B topo	0.260	1.828	0.488	0.774
LLaMA3.2 1B SFT	0.219	1.322	0.562	0.653
LLaMA3.2 1B topo	0.266	1.228	0.589	0.594
LLaMA3.2 3B SFT	0.401	1.000	0.577	0.638
LLaMA3.2 3B topo	0.395	0.962	0.631	0.566
Qwen2.5 7B SFT	0.408	0.879	0.656	0.520
Qwen2.5 7B topo	0.463	0.802	0.667	0.504

Table 3: Results for edge-level F_1 and normalized graph edit distance (nGED). Gemma3 27B and Qwen3 235B use in-context learning, the rest supervised fine-tuning.

Supervised fine-tuning. Table 2 shows that multi-task training with topological sorting as an explicit predicting objective yields stable improvements over single-task ARI training. These results suggest that explicit topological supervision is a useful inductive bias, leading to improved performance.

5.1.2 Structural results

To examine why topological (topo) sorting improves model performance on the ARI task, we analyze both edge-level F_1 which reflects the local structure of the argument graph and normalized graph edit distance which captures global structural similarity (see Section 3.2). Table 3 shows results for in-context learning and supervised fine-tuning. We can see that incorporating topological sorting consistently improves edge-level F_1 while reducing nGED. This indicates that topological reasoning helps models better capture global graph structure, leading to more accurate local relation predictions and improved overall ARI performance.

5.2 Ablation and Analysis

We analyze why topological reasoning improves ARI under two settings, in-context learning (ICL) and supervised fine-tuning (SFT), examining

whether gains result from extra decoding, generic CoT-style reasoning, implicit structural reasoning without committing to an order, or explicit, graph-aware topological ordering.

In-context learning ablations. We compare several prompt variants that differ only in the reasoning signal provided at inference: *relation only* outputs only the relation list; *relation-topo* predicts the relation list then outputs the topological order; *topo-relation* outputs the topological order then the relation list; *random order* is similar to relation-topo but replaces the topological order with a random order (testing if the gains are due to extra decoding); *hidden topo* reasons about the order internally but outputs only the relation list (testing if implicit structure suffices); *CoT* outputs reasoning steps before the argument relations list (testing if gains are from general reasoning rather than explicit topological reasoning). We evaluate Gemma 3 27B (Table 4) and LLaMA 3.3 70B (Table 18 in Appendix E) on AbstRCT and CDCP.

Our key findings are: (1) Graph-aware ordering, not "just more decoding": random order matches the decoding budget and format of topological prompting but does not improve ARI or structural metrics, ruling out extra tokens or steps as the source of gains. (2) Structural graph reasoning, not "just CoT": CoT does not improve ARI, indicating that improvements are driven by graph-structured global constraints rather than generic linear reasoning. (3) Explicit commitment outperforms implicit reasoning: when the model reasons about order but not outputs it (hidden topo), performance reverts to the relation only prompt, suggesting that committing to a global order constrains relation decoding and enforces coherence. (4) Order placement (relation-topo vs topo-relation): both outperform the relation only prompt, with a slight advantage for relation-topo, indicating that predicting the topological order after generating the local relations provides a stronger global structural signal.

Overall, in-context learning ablations indicate that improvements in ARI and structural metrics are driven by graph-aware topological ordering, not by extra decoding, generic CoT, or hidden/implicit structure. EO and GO are high when the order is predicted explicitly, confirming internal coherence and alignment to the gold partial order.

Supervised fine-tuning ablations. As shown in Table 2, incorporating topological ordering during training improves ARI beyond standard fine-tuning. To test whether this contribution stems from struc-

Prompt	AbstRCT					CDCP				
	ARI F1	edge F1	nGED	EO	GO	ARI F1	edge F1	nGED	EO	GO
relation only	0.655	0.384	1.82	-	-	0.542	0.146	2.324	-	-
relation-topo	0.719	0.488	1.243	0.978	0.882	0.563	0.192	2.051	0.858	0.516
topo-relation	0.679	0.419	1.812	0.969	0.880	0.558	0.179	2.166	0.851	0.478
random order	0.664	0.390	1.561	0.592	0.661	0.542	0.144	2.255	0.487	0.508
hidden topo	0.651	0.374	1.802	-	-	0.546	0.157	2.722	-	-
CoT	0.650	0.369	1.683	-	-	0.550	0.167	2.556	-	-

Table 4: Results of different prompts for Gemma 3 27B. EO: edge-order consistency, GO: gold-order agreement.

Model	CDCP		PE	
	random	topo	random	topo
LLaMA3.2 1B	0.597	0.614	0.757	0.779
Qwen2.5 1.5B	0.557	0.585	0.783	0.791

Table 5: Comparison between random order and topological order in the auxiliary task.

tural information rather than additional supervision, we compare two training objectives with identical format and supervision cost: topological supervision (which predicts a valid topological order of the gold argument graph) and random supervision (which uses a random permutation, serving as a structure-neutral negative control). Table 5 shows that replacing gold topological orders with random permutations during training consistently degrades ARI F_1 across models and datasets. This indicates that the observed gains come from graph-aware ordering rather than additional supervision signals.

Summary. Across in-context learning and supervised fine-tuning, only conditions with a meaningful graph-aware topological order improve both local (edge-level F_1) and global (nGED, EO, GO) structure. We conclude that explicit global ordering provides a model-agnostic mechanism for injecting structural coherence into ARI, complementing both prompting and supervised training.

6 Structural Faithfulness and Verifiable Reasoning

Beyond improving ARI performance, our goal is to provide faithful and verifiable explanations of model behavior. Unlike post-hoc natural-language rationales, our method exposes the model’s global structural reasoning through an explicit topological ordering over argument components.

Next, we define the notion of explanation in this setting and provide quantitative and qualitative evidence that the predicted topological order acts as a faithful mediating mechanism for ARI predictions.

Model	CDCP		PE	
	EO	GO	EO	GO
Gemma3 27B topo	0.858	0.516	0.908	0.519
Qwen3 235B topo	0.808	0.615	0.925	0.754
LLaMA3.2 1B topo	0.678	0.664	0.854	0.820
LLaMA3.2 3B topo	0.886	0.822	0.869	0.814
Qwen2.5 7B topo	0.941	0.911	0.920	0.841

Table 6: Edge-order consistency and gold-order agreement for in-context learning (Gemma3 27B and Qwen3 235B) and supervised fine-tuning (rest of the models).

6.1 What Constitutes an Explanation in ARI

An explanation for ARI is a model-generated intermediate representation that (i) is structurally grounded, (ii) is internally consistent with the model’s predictions, and (iii) allows independent verification against gold argumentative structure.

Under this definition, the predicted topological order qualifies as an explanation as it encodes a partial order over components, constraining admissible relations in the directed acyclic argument graph. Crucially, the predicted order is not computed post-hoc, it is generated by the model as part of inference or training objective, and thus reflects its internal global reasoning process.

6.2 Internal Faithfulness & External Validity

We evaluate the quality of the model’s reasoning traces using two complementary, order-based metrics introduced in Section 3.2: EO measures *internal faithfulness* (whether predicted edges are consistent with the model’s own predicted order), GO measures *external validity* (how well the predicted order aligns with the gold argumentative flow). Table 6 shows consistently high EO scores (often exceeding 0.85), demonstrating strong internal coherence between predicted orders and predicted relations. At the same time, high GO values show that the model’s inferred reasoning paths align closely with the gold argument structure. Together, these results indicate that explicit topological reasoning

Bucket	F_1	nGED	reverse rate
Low-GO	0.091	1.403	0.417
Medium-GO	0.405	0.970	0.221
High-GO	0.545	0.867	0.051

Table 7: Causal mediation on PE dataset for Gemma-3 27B with topological prompting.

Setting	F_1	nGED
Baseline (Topo)	0.339	1.001
Adjacent swap	0.144	1.487
Sink displacement	0.145	1.496

Table 8: Counterfactual order perturbation on PE dataset using Gemma-3 27B.

yields explanations that are both faithful to the model’s decision process and structurally consistent with human-annotated argumentative flow.

6.3 Causal Mediation Analysis

We test whether the improvements from topological prompting are mediated by the correctness of the model-generated global order. Specifically, we treat the use of topological prompting as the intervention X , ARI performance as the outcome Y , and the quality of the predicted order as the mediator M , measured by GO. We conduct this analysis on the PE dataset with Gemma3 27B. We stratify model outputs by GO into tertiles (Low: $GO < 0.25$; Medium: $0.25 \leq GO < 0.822$; High: $GO \geq 0.822$) so that each bucket has roughly the same number of examples. The results are shown in Table 7. We can see that ARI performance improves monotonically with GO, while direction-reversal rate (the percentage of predicted edges that reverse the direction of the ground truth edges) and nGED correspondingly decrease monotonically. These monotonic trends indicate that performance gains arise primarily when the predicted order is structurally correct, providing evidence that the topological order functions as a faithful mediating mechanism for argument relation identification rather than a post-hoc explanation.

6.4 Counterfactual Order Perturbation

To further test whether the predicated global order acts as a causal reasoning mechanism, we perform counterfactual interventions on the mediator by perturbing the predicted topological order while keeping the input text fixed. We apply controlled order modifications: *adjacent swaps of components* (swapping a middle adjacent pair) and *sink*

Error Type	Num errors	Fixed by topo	Fix rate given error
direction reversals	163	49	30.06%
spurious chains	127	14	11.02%
missing supports	51	23	45.10%

Table 9: Error analysis on PE dataset, using Gemma3 27B with in-context learning.

Order source	sink-late (k=1)	sink-late (k=2)	Q4 target-ing
random order	0.223	0.451	0.315
predicted order	0.337	0.535	0.531
gold order	0.985	0.990	0.896

Table 10: Structural read-off analysis for Gemma3 27B on PE dataset.

displacement (moving the last node, typically the claim/sink, to the front), and then re-decode argument relations conditioned on the perturbed order.

Table 8 shows that, across the PE dataset, these interventions lead to systematic degradations in structural performance. Importantly, these changes are structured rather than random, indicating that relation predictions are causally dependent on the ordering. This intervention-based test further supports the claim that the topological order functions as a mechanistic driver of ARI predictions.

6.5 Diagnosing Common ARI Errors

We next assess if reasoning traces help analyze and correct common ARI failures. We consider three error types: *direction reversals* when the model predicts $v \rightarrow u$ instead of the gold edge $u \rightarrow v$ i.e. claim to premise instead of premise to claim; *spurious premise chains* (premises linked serially rather than in parallel to a claim), gold shows ≥ 2 supports ($s_i \rightarrow c$) to a sink c (in-degree > 0 , out-degree $= 0$), but the prediction reduces the number of ($s_i \rightarrow c$) edges and adds chain edges among the supporters ($s_i \rightarrow s_j$); *missing supports to a common sink* (omitted edges to the main claim), gold shows ≥ 2 supports to a sink c , but the prediction misses one or more ($s_i \rightarrow c$) edges without adding chains among supporters.

We conduct an error analysis on Gemma3 27B for PE dataset using the AR prompt (see baseline in Section 4) and topological prompt. Error types are computed for each paragraph and a fix is counted when an error appears with the AR prompt but disappears under the topological prompt for the same paragraph. Table 9 shows that topological prompting reduces directionality errors (fixing $\sim 30\%$ of AR prompt cases) and is particularly effective at

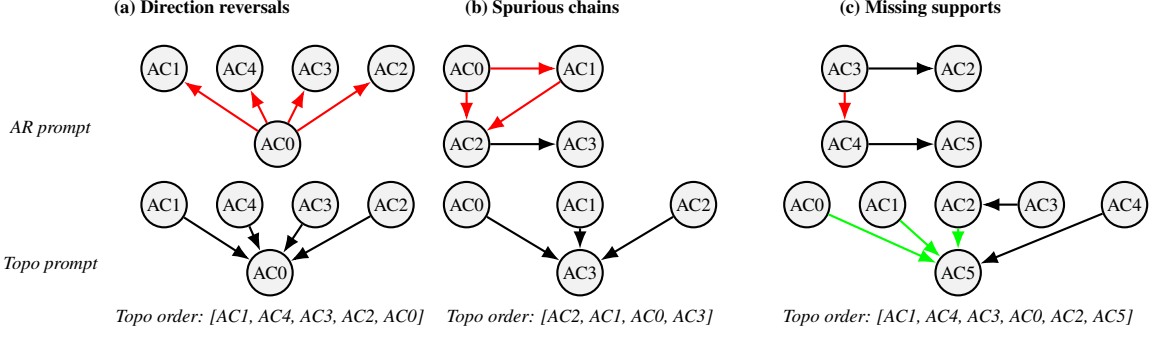


Figure 3: Examples of ARI error types. The top graphs show the AR-prompt (see baseline in Section 4) prediction, the lower graphs show the topological-prompt predictions and its predicted topological order. Red edges are wrong predictions, green are ground-truth edges missed by the AR prompt. Topo order (a) enforces claim-last behavior, fixing reversed edges, (b) removes spurious chains and all supports point to the sink, (c) restores all supports.

Prompt type	F1	edge F1	nGED
prediction only	0.721	0.473	0.100
Adding topological sort	0.741	0.513	0.960

Table 11: Results for discourse relation parsing.

recovering missing supports to a common sink (45.10% fix rate). Figure 3 shows how explicit topological ordering corrects ARI errors.

6.6 Structural Read-off from Predicted Topological Orders

We evaluate if the model’s predicted topological orderings encode recoverable global argumentative structure using a structural read-off analysis. In particular, we assess whether sinks (claims) appear late in the predicted order and whether argumentative relations preferentially target late-order nodes.

We compare three order sources: random order (chance baseline), predicted order produced by the model, and gold topological order (averaged over multiple valid topological sorts). For each paragraph with a non-empty topological order, we compute two order-based metrics: *sink-late placement*, the fraction of gold sink nodes appearing in the final k positions of the order ($k = 1, 2$); and *sink preference in edge direction (Q4 targeting)*, the fraction of predicted edges whose target nodes are in the last quartile of the order. These metrics quantify whether global argumentative functions are recoverable directly from the ordering. Table 10 shows that predicted order outperforms random baseline, placing the gold sink last more frequently and directing more relations towards late-order nodes. When normalized between random and gold (computed as $(M_{pred} - M_{rand}) / (M_{gold} - M_{rand})$), predicted orders recover $\sim 15\text{--}16\%$ of the signal for sink-late

placement and 37% for Q4 targeting. While exact sink positioning remains challenging,⁶ these results indicate that the model learns a coherent claim-late, sink-directed ordering bias, supporting explicit topological ordering as a meaningful and interpretable intermediate representation.

7 Explicit Topological Reasoning for Discourse Relation Parsing

To assess the generality of explicit topological reasoning, we apply our method to discourse relation parsing, where relations among EDUs induce a DAG structure. We use SciDTB (Yang and Li, 2018) and Gemma3 27B with in-context learning. We replace argumentative components with discourse units and predict a topological order alongside discourse relations. Similarly to ARI, we ask the model to output directed, unlabeled relation pairs and evaluate macro F_1 for relation pairs, and edge F_1 and nGED for structural metrics. Table 11 shows consistent improvements in relation F_1 and structural metrics, mirroring the trends observed in ARI. This suggests that explicit global ordering serves as a task-agnostic mechanism for enforcing coherence in structured prediction problems.

8 Conclusion

We reframed ARI as a global structural reasoning problem and introduced topological ordering as an explicit prediction target to make global structure explicit. Our method is model-agnostic, applicable to different models and learning paradigms. Extensive experiments show that explicit topological reasoning improves accuracy, structural fidelity

⁶Metrics increase consistently with model size, suggesting that the recoverable ordering signal is capacity-dependent.

and explainability in ARI. We demonstrate the generality of our method on discourse relation parsing, showing consistent improvement. In future work, we plan to extend our method to other graph-structured prediction tasks with LLMs.

9 Limitations

While topological ordering improves structural reasoning, the approach assumes that the underlying argument graph can be reasonably approximated as a directed acyclic graph. Argumentation in real-world texts may involve cycles, implicit relations, or underspecified structures that are not fully captured by this assumption. In addition, for in-context learning, the effectiveness of topological prompting depends on the model’s ability to follow structured output instructions, weaker or smaller models may benefit less. Moreover, our evaluation focuses on benchmark datasets with relatively clean annotations, and performance on noisier argumentation remains to be explored. Finally, because large language models are trained on broad web-scale corpora, it is possible that common ARI datasets (e.g., AbstRCT, CDCP, PE) appear in their pretraining data. However, all variants in our study use the same underlying pretrained model, so any such exposure would affect them equally and cannot explain the relative improvements from topological ordering. We acknowledge this as a general limitation of LLM-based evaluation.

References

- Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. [A neural transition-based model for argumentation mining](#). 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 6354–6364, Online. Association for Computational Linguistics.
- Jianzhu Bao, Yuhang He, Yang Sun, Bin Liang, Jiachen Du, Bing Qin, Min Yang, and Ruifeng Xu. 2022. [A generative model for end-to-end argument mining with reconstructed positional encoding and constrained pointer mechanism](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10437–10449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Maciej Besta, Florim Memedi, Zhenyu Zhang, Robert Gerstenberger, Guangyuan Piao, Nils Blach, Piotr Nyczyk, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, and 1 others. 2025. Demystifying chains, trees, and graphs of thoughts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- 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.
- Lucas Carstens and Francesca Toni. 2015. [Towards relation based argumentation mining](#). In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics.
- 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.
- Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a graph: Encoding graphs for large language models.
- 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.
- Haoyu Han, Yaochen Xie, Hui Liu, Xianfeng Tang, Sreyashi Nag, William Headden, Yang Li, Chen Luo, Shuiwang Ji, Qi He, and Jiliang Tang. 2025. [Reasoning with graphs: Structuring implicit knowledge to enhance LLMs reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25698–25714, Vienna, Austria. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*.
- John Lawrence and Chris Reed. 2019. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- Yihao Li, Ru Zhang, and Jianyi Liu. 2024. An enhanced prompt-based llm reasoning scheme via knowledge graph-integrated collaboration. In *International Conference on Artificial Neural Networks*, pages 251–265. Springer.
- Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, and Sophia Ananiadou. 2023. [Argument mining as a multi-hop generative machine reading comprehension task](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10846–10858, Singapore. Association for Computational Linguistics.

- L Luo, YF Li, G Haffari, and S Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *ICLR 2024: The Twelfth International Conference on Learning Representations*. ICLR.
- Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.
- Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, and Kohsuke Yanai. 2022. [End-to-end argument mining with cross-corpora multi-task learning](#). *Transactions of the Association for Computational Linguistics*, 10:639–658.
- 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.
- 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)*.
- Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(6):62–70.
- Christian Stab and Iryna Gurevych. 2014. [Identifying argumentative discourse structures in persuasive essays](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. 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, 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.
- An Yang and Sujian Li. 2018. [SciDTB: Discourse dependency TreeBank for scientific abstracts](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

A Prompt

Prompt for our proposed method:

You are an expert in Argument Mining. You are given a biomedical abstract containing numbered argument components enclosed in <AC> tags (e.g., <AC1>...</AC1>).

Your task has two parts:

1. Identify argument relations between argument components in the abstract text.

2. Construct the reasoning path

Assume the argument structure forms a directed acyclic graph (DAG), where edges represent argumentative relations (e.g., support or attack).

Return a topological sort of the graph that reflects the reasoning path from foundational premises to final claims.

Output Format

Return a single JSON object with two keys:

```
{
  "argument_relations_list": [["source AC",
    "target AC"],..., ["source AC", "target AC"]],
  "reasoning_path": [
    "<AC3>",
    "<AC2>",
    "<AC1>",
    ...
  ]
}
```

"source AC" and "target AC" are argument components (ACs) in the abstract text (such as "<AC1>", "<AC2>", ...).

"reasoning_path": a list of <AC> tags in topological order, from foundational premises to final claims.

Some paragraphs do not contain any argument components (without <AC> tags), for these paragraphs, output an empty list for "reasoning_path" and "argument_relations_list". Do not include any explanations or extra text, just the JSON object.

Prompt for the relation only baseline (**ari**):

You are an expert in Argument Mining. You are given a biomedical abstract containing numbered argument components enclosed in <AC> tags (e.g., <AC1>...</AC1>).

Your task is to identify argument relations between argument components in the abstract text.

Output Format

Return a single json object containing the argument component pairs:

```
{"argument_relations_list": [["source AC",
  "target AC"],..., ["source AC", "target AC"]]}
"source AC" and "target AC" is argument components (ACs) in the essay text (such as "<AC1>", "<AC2>", ...).
```

Do not include any explanations or extra text, just the json object.

Prompt for the CoT baseline:

You are an expert in Argument Mining. You are given a biomedical abstract containing numbered argument components enclosed in <AC> tags (e.g., <AC1>...</AC1>).

Your task is to identify argument relations between argument components in the abstract text. Let's think step by step.

Output Format

Return a single JSON object with two keys:

```
{
  "reasoning_path": Your reasoning process,
  "argument_relations_list": [["source AC",
    "target AC"],..., ["source AC", "target AC"]]
}
```

"reasoning_path": Your reasoning process.
"source AC" and "target AC" is argument components (ACs) in the text (such as "<AC1>", "<AC2>", ...).
Do not include any explanations or extra text, just the json object.

Prompt for the pairwise baseline:

You are an expert in Argument Relation Identification (ARI). Determine whether the first argument component (SOURCE) has an argumentative relation with the second argument component (TARGET).

Rules:

- Output only a single digit:
- 1 → if SOURCE has an argumentative relation with TARGET
- 0 → if there is no relation between SOURCE and TARGET
- Do not output anything else.

SOURCE:

"A beneficial effect of MPA on appetite was observed after both 6 weeks (P = .008) and 12 weeks (P = .01) of treatment."

TARGET:

"After 12 weeks, a mean weight gain of 0.6 +/- 4.4 kg was seen in the MPA, versus an ongoing mean weight loss of 1.4 +/- 4.6 kg in the placebo group."

Label:

Prompt for topo_relation in the ablation studies:

You are an expert in Argument Mining. You are given a biomedical abstract containing numbered argument components enclosed in <AC> tags (e.g., <AC1>...</AC1>).

Your task has two parts:

1. Construct the reasoning path
Assume the argument structure forms a directed acyclic graph (DAG), where edges represent argumentative relations (e.g., support or attack).
Return a topological sort of the graph that reflects the reasoning path from foundational premises to final claims.
2. Identify argument relations between argument components in the abstract text.

Output Format

Return a single JSON object with two keys:

```
{
  "reasoning_path": [
    "<AC3>",
    "<AC2>",
    "<AC1>",
    ...
  ],
  "argument_relations_list": [["source AC",
    "target AC"],..., ["source AC", "target AC"]]
}
```

"reasoning_path": a list of <AC> tags in topological order, from foundational premises to final claims.
"source AC" and "target AC" are argument components (ACs) in the abstract text (such as "<AC1>", "<AC2>", ...).

Some paragraphs do not contain any argument components (without <AC> tags), for these paragraphs, output an empty list for "reasoning_path" and "argument_relations_list".
Do not include any explanations or extra text, just the JSON object.

Prompt for random_order in the ablation studies:

You are an expert in Argument Mining. You are given a biomedical abstract containing numbered argument components enclosed in <AC> tags (e.g., <AC1>...</AC1>).

Your task has two parts:

1. Identify argument relations between argument components in the abstract text.
2. Provide a random order of the components.

Output Format

Return a single JSON object with two keys:

```
{
  "argument_relations_list": [["source AC",
    "target AC"],..., ["source AC", "target AC"]],
  "reasoning_path": [
    "<AC3>",
    "<AC2>",
    "<AC1>",
    ...
  ]
}
```

"source AC" and "target AC" are argument components (ACs) in the abstract text (such as "<AC1>", "<AC2>", ...).
"reasoning_path": a list of <AC> tags in random order.

Some paragraphs do not contain any argument components (without <AC> tags), for these paragraphs, output an empty list for "reasoning_path" and "argument_relations_list".
Do not include any explanations or extra text, just the JSON object.

Prompt for hidden topo in the ablation studies:

You are an expert in Argument Mining. You are given a biomedical abstract containing numbered argument components enclosed in <AC> tags (e.g., <AC1>...</AC1>). Your task is to identify argument relations between argument components in the abstract text.

Before answering, think about the logical flow and determine a topological order internally, but do not include the order in your final answer.

Output Format

Return a single json object containing the argument component pairs:

```
{"argument_relations_list": [{"source AC", "target AC"}, ..., [{"source AC", "target AC"}]}
```

"source AC" and "target AC" are argument components (ACs) in the abstract text (such as "<AC1>", "<AC2>", ...).

Do not include any explanations or extra text, just the JSON object.

B Dataset licenses and statistics

We use three public available datasets for our experiments: AbstRCT (Mayer et al., 2020), CDCP (Park and Cardie, 2018) and PE (Stab and Gurevych, 2014). The licenses of these datasets are:

AbstRCT: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

CDCP: ODC Open Database License v1.0 ^{7,8}.

Persuasive Essays: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0).

All datasets are in English.

Table 12, Table 13 and Table 14 shows the train test split for these datasets.

split	num
train	350
gla test	100
mix test	100
neo test	100

Table 12: Train test split for abstRCT dataset.

split	num
train	580
test	150

Table 13: Train test split for CDCP dataset.

split	num
train	1796
test	439

Table 14: Train test split for PE dataset.

Hyperparameter	Value
r(rank)	8
lora_alpha	16
bias	none
target_module	all_linear

Table 15: Hyperparameters for PEFT.

C Fine-tuning and prompting configurations

We use HuggingFace SFTTrainer and PEFT ⁹ for all the experiments. The hyperparameters for PEFT is shown in Table 15. The training hyperparameters are reported in Table 16. For generation, we use do_sample = False.

Hyperparameter	value
max_length	1024
per_device_train_batch_size	8
gradient_accumulation_steps	1
num_train_epochs	5
lr_scheduler_type	cosine
warmup_ratio	0.1
learning_rate	1e-4
bf16	True

Table 16: Training hyperparameters used in the experiments.

For the prompting experiments, we use greedy decoding for all the experiments, temperature = 0.0, top_p = 1.0. Gemma models are accessed using Google AI studio, all the other API-based experiments were conducted through the OpenRouter platform (<https://openrouter.ai>), which provides unified access to multiple model providers. The model, provider and identifier are shown in Table 17.

D Macro F1 score for ARI task

For the Argument Relation Identification (ARI) task, the model outputs ARI predictions in the following format:

⁷<http://www.opendatacommons.org/licenses/odbl/1.0/>

⁸<https://spdx.org/licenses/ODbL-1.0.html>

⁹https://huggingface.co/docs/trl/v0.26.1/en/sft_trainer#trl.SFTTrainer

Model	Provider	Identifier
Gemma 3 12B	Google AI studio	gemma-3-12b-it
Gemma 3 27B	Google AI studio	gemma-3-27b-it
LLaMA 3.3 70B	deepinfra/turbo	meta-llama/llama-3.3-70b-instruct
Qwen 3 235B	deepinfra/fp8	qwen/qwen3-235b-a22b-2507
GPT 4.1	openai	openai/gpt-4.1

Table 17: Provider and identifier for the models we use in the prompting experiments.

```
{
  "argument_relations_list": [
    ["source AC", "target AC"],
    ...,
    ["source AC", "target AC"]
  ]
}
```

During evaluation, we obtain a list of dictionaries, each corresponding to a paragraph in the test set. Each dictionary contains two keys: `predict` (the model’s predicted relations) and `label` (the gold relations), both following the format above. To compute the F1 score for ARI, we follow these steps:

1. **Initialization.** We create two lists, `gt_list` and `pred_list`, which record whether a specific pair of argument components is connected by a relation. Each entry is either `relation` (indicating the presence of a relation) or `none` (indicating no relation).
2. **Enumerating all possible component pairs.** For each paragraph, we determine the maximum argument component index, denoted as `max_ac`. Then, for both the predicted and gold `argument_relations_list`, we examine every ordered pair (i, j) where $0 \leq i, j \leq \text{max_ac}$. If the pair $[\langle AC_i \rangle, \langle AC_j \rangle]$ appears in the predicted or gold list, we append `relation` to `pred_list` or `gt_list`, respectively; otherwise, we append `none`.
3. **Metric computation.** We compute the macro F1 score over the full `pred_list` and `gt_list`.

E Additional ablation study results

Prompt	AbstRCT					CDCP				
	ARI F1	edge F1	nGED	EO	GO	ARI F1	edge F1	nGED	EO	GO
simple prompt	0.642	0.356	1.771	-	-	0.533	0.133	2.547	-	-
relation-topo	0.691	0.440	1.458	0.994	0.902	0.549	0.163	2.314	0.883	0.394
topo-relation	0.678	0.416	1.494	0.989	0.896	0.554	0.174	2.308	0.940	0.450
random order	0.603	0.277	1.792	0.444	0.415	0.536	0.135	2.381	0.473	0.558
hidden topo	0.648	0.368	1.814	-	-	0.533	0.126	2.901		
CoT	0.648	0.365	1.785	-	-	0.532	0.125	2.852	-	-

Table 18: Experiments of different prompts for LLaMA 3.3 70B. EO: edge-order consistency, GO: gold-order agreement.